THREE ESSAYS ON CONSUMERS' ACTIVITIES IN THE ONLINE DOMAIN

by

Shaoqiong Zhao

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in

Management Science

at

The University of Wisconsin-Milwaukee

July 2014

ABSTRACT

THREE ESSAYS ON CONSUMERS' ACTIVITIES IN THE ONLINE DOMAIN

by

Shaoqiong Zhao

The University of Wisconsin-Milwaukee, 2014

Under the Supervision of Dr. Sanjoy Ghose

Nowadays, with the explosive growth in the usage of the Internet, consumers are performing all kinds of activities over the Internet like searching or buying. We want to study the different activities of consumers in the online domain.

In our daily lives, people are often making various kinds of product purchases. When making such purchases, a lot of factors can affect consumers' decisions. This includes the nature of the product category, and especially in the online domain, the nature of their search activities. In the first essay/chapter, we develop an econometric model to understand the relationships between different dimensions of on-line search and purchase behavior. Our approach uses endogeneity corrections to develop a model that is more correct than the typical non-endogeneity corrected model. Thus we believe our results to be truly reflective of what is happening in the search-buying domain. We use extensive empirical data to test several hypotheses that we developed. Parameters from our model

estimations reveal that there are interesting variations in the search-purchase behavior relationships across types of product categories. This difference is especially evident between utilitarian and hedonic goods. Our findings have important theoretical and managerial implications.

The amount of information in text reviews is tremendously greater than that in typical numerical data. A major challenge for marketers is how to extract the most relevant information from this big data source. In our second essay/chapter, we do this by using a text mining methodology that draws on machine learning algorithms. We collect data using a Java WebCrawler type programming approach. We use a word-based model to predict consumers' recommendations. Model prediction accuracy was high. In the marketing literature there has been almost no work where such a methodology has been used to make predictions of recommendations based on big data stemming from textual information. An interesting finding from our research is that as the number of textual features increases, the predictive accuracy of the model increases only up to a point. Beyond that, inclusion of more words in the model leads to a decrease in predictive accuracy. We also use a diagnostic approach to identify key words that are determinants of user recommendations. Since our model deals with big data, we address in details the issue of scalability; our computations show that our approach is very scalable. Potential for marketing implications seems considerable.

Marketers are always interested in predicting market sales so that they can arrange the firm activities accordingly. In the meantime, this market sales information can also help the consumers to make right buying decisions. However the high cost and long period of

collecting the available data with a lag makes it very inconvenient and out of date. With the rise of multi-social media sharing websites such as YouTube, Flickr, and various blogs, consumers can search and learn various types of information from these websites. The availability of large amounts of data on the Internet enables us to use large scale data mining algorithms for solving complex problems. The users' online searching activities can be captured for predicting the market sales. In the third essay/chapter, we focus on the impacts of different search behavior and marketing outcomes like product sales. We examined the three major online search areas including text, image, and video from search engines like Google to help us accurately and easily predict the sales of automobiles. We believe that our work here opens a brand new arena for using multimedia search activities and will have a big impact on marketing sciences.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

This body of work is a milestone in my life.

First, I would like to express my deepest gratitude to my dear major Professor and advisor, Dr. Sanjoy Ghose. Throughout the time of my Ph.D. study at UWM, Dr. Ghose provided invaluable guidance and tremendous support. His endless pursuits in academic research inspired me to continue on this professional endeavor. It is my great honor to be his advisee and I really appreciate the opportunity to learn from him and work with him.

I would also like to thank Dr. V. Kanti Prasad, Dr. Amit Bhatnagar, Dr. Xiaojing Yang and Dr. Tingting He for taking the time to serve on my dissertation committee, and for their extensive discussions and suggestions on this work. I am also grateful of Dr. Ehsan Soofi and Dr. Timothy Haas; they helped me with lots of statistical problems.

Lastly, and most importantly, I wish to thank my parents, Minggen Zhao and Yuee Guo, who have supported me with their endless love and care. This research would not have been possible without them.

Chapter/Essay 1

Identifying Relationships between Online Buying and Online Search Behaviors

1.1. Introduction

For the past few decades, researchers and marketing managers have been interested in understanding what can affect consumer purchase habits. Just like Dhar and Wertenbroch in their 2000 paper said that, consumer choices are driven by utilitarian and hedonic considerations. Consumer attitudes have distinct hedonic and utilitarian components and products differ in the extent to which their overall attitudes are derived from the two components. These different perspectives of consideration let people distinguish goods between hedonic and utilitarian nature and make decisions according to their preference (Batra & Ahtola, 1990; Mano & Oliver, 1993). Broadly speaking, the term hedonic refers to aesthetic, experiential, and fun benefits (Dhar & Wertenbroch, 2000; Strahilevitz & Myers, 1998), and the utilitarian term refers to the functional, instrumental and practical benefits (Hirschman & Holbrook, 1982). However, hedonic and utilitarian components are not two ends of a one-dimensional scale (Okada, 2005); products vary in the perceived level of hedonic or utilitarian components (Batra & Ahtola, 1990). We can classify a product as a mainly hedonic one if the hedonic components make up the major part of the product as in say, movies. A movie review by Holden in 2005 from The New York Times demonstrated movie possesses a strong hedonic feature. Some other products like tools will be viewed as a utilitarian good since it is primarily more utilitarian and its main benefits to the consumers are the functions associated with the consumption. There are still other products, which cannot be simply classified as a hedonic or utilitarian one

due to the co-occurrence of both aspects. Both hedonic and utilitarian components express an equal weight in the benefits offered by the products. In apparels for instance, people try to get their basic needs fulfilled—the need to be covered. That is a totally utilitarian aspect. At the same time, people would like to express a certain self-image by dressing in a certain style of clothes, and this is a hedonic aspect. When facing different purchase situations, whether it is a hedonic, or utilitarian, or mixed product, the behavior will be quite different.

With the explosive growth of e-commerce activities, people nowadays are making more and more purchases online. The Internet and the World Wide Web (WWW or the Web) in particular, represents a recent technological innovation that has had a profound impact on all facets of people's lives" (Lin & Yu, 2006). When consumers are making purchases they get to put effort into this activity. According to Andreasen (1968) there are five major types of information sources including Impersonal Advocate, Impersonal Independent, Personal advocate, Personal Independent, and direct observation. It is not easy to measure consumers' search efforts for these five types of information, and most of the time this was done by taking survives of the customers. However when we put this into the online domain it becomes quite simple and we can easily record consumer search behavior directly using technology. Also the idea that consumers differ in the amount and type of effort they put into shopping like searching effort is not new to marketing (Katona & Mueller, 1955, Newman & Staelin, 1972). Especially when it comes to the search of information at e-commerce context, it is different from traditional search in various ways like information source, type, etc. Such differences are important to marketers because

they influence consumers' reactions to marketing strategies. And these different perspectives of consideration also let people fall into different purchasing/searching patterns based on their own individual characteristics like demographic characteristics and the external product characteristics like product categories.

Most of the previous research of consumers' buying behavior had focused on simply the choice across different brands made by the consumers. There is a great need to look at how search effort and other factors could impact the more detailed purchase behavior like how much consumers are willing to pay for each item. At the same time, we would also want to explore how by nature the willingness to pay could impact the consumers search behavior also. To fill this gap and explore the truth, the current research proposes a simultaneous model by digging into the interdependence of the search effort and willingness to pay across two distinct product categories: hedonic and utilitarian goods. The main contribution of this study is to quantitatively and comprehensively analyze the relationships between search effort and willingness to pay under the online domain and empirically aggregate and generalize the results across the hedonic and utilitarian product categories. By using simultaneous modeling, we investigate the inter-impact of the two major online consumer activities: search behaviors and purchase behaviors, and also how the impacts vary across the product categories of hedonic and utilitarian. We consider the context of E-commerce and provide quantitative generalizations on 4 product categories (2 of hedonic, 2 of utilitarian) over 10,000 consumers through the whole 2004 calendar year. Using the three-stage-least-square estimation methods, we systematically integrate and uncover: 1) the interdependence between search effort and willingness to pay of

consumers made online and 2) the moderating effect of product categories on the interdependence.

The remainder of the article is organized as follows. We next provide the theoretical background from detailed literature review for our research. We then discuss the methodology that we use for the study. Then we present the data as well as the empirical approaches and results. We conclude with summarizing our findings, a discussion of the theoretical and managerial implications and future research directions.

1.2. Theoretical Framework and Hypotheses

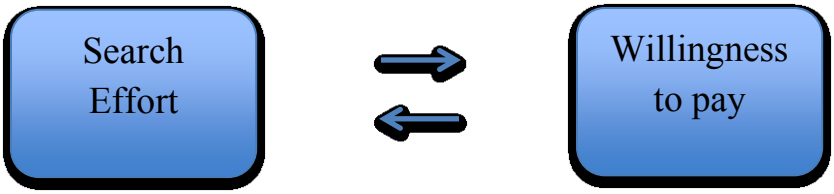
In previous empirical studies, researchers examined the impact of search effort on purchase choice (Chaney, 2000). This search effort refers to the traditional search effort but it is somewhat similar to search effort under the online domain. So we first review how search effort impact the consumers purchase behavior and how the willingness to pay would in turn impact the search effort. Then we will review how demographics impact consumers search effort and purchase behavior. After this we would like to define the terms of “hedonic”, “utilitarian” and provide a brief review of the prior research relevant to our study.

1.2.1. Search Effort and Purchase Behavior

Involvement is a psychological construct, which was pioneered by Sherif and Cantril (1947), who described involvement as the state of an organism when presented with any stimulus, which is ego central, or when any stimulus is either consciously or

subconsciously related to the ego. Involvement with something ordinarily influences attitudes and behaviors relating to it. Thus involvement with purchasing would influence attitudes and behavior relating to it. In real life, consumers perform online search before they actually conduct the purchase online. However, in the mental process of this whole buying behavior, searching and purchasing happen simultaneously and initiating of either one behavior gets the consumers involved in the big buying decision process. We argue that when a consumer starts a search on the Internet, he/she gets involved in this online shopping activity; so his/her following purchasing behavior like purchase choice or expenditure will be impacted by the search activity. Vice versa, when the consumer has considered spending certain amount of money, he/she gets involved with this intending purchase, and this will impact how he/she will perform the online search activity to help him/her make the final purchase decisions. The searching and buying are happening simultaneously in the consumers' mental process and impact each other. In our study, we inspected how the willingness to pay and search effort interacts with each other. So we come up our first hypothesis.

Hypothesis 1: There is interdependence between online search effort and willingness to pay.



Kassarjian (1981) has recently related search effort to the notion of consumer involvement and proposed that consumers' involvement with purchasing influences their

purchase behavior. From the consumers' purchasing behavior perspective, information plays an important role in it. But where do the consumers get the information? In the traditional offline situation, consumers can look for the information from five types of information source (Andreansen, 1968). These are the impersonal advocate, like mass media advertising; impersonal independent, like consumers reports; personal advocate, like sales people's advice; personal independent, like opinions of co-workers; and the last direct observation, like a trail or demonstration.

The Internet is a useful tool for information search (Hammond, Mcwilliams & Diaz, 1998). Internet makes a large volume and variety of information available so that consumers can easily acquire information from web sites that is similar to the information available from traditional mass-media advertising (Peterson& Merino, 2003). So when it comes to the e-commerce context, the search effort mainly reflects the impersonal advocate, which is a public information source where consumers can identify the relevant information, just like reading a magazine, or watch television commercials. Information search involves both cognitive and physical effort (Johnson, Bellman, & Lohse, 2003). For these researchers, the extent of information search, often measured by the number of acquisitions or the number of products viewed, occurs within, not just across, retailers and other providers of information (Diehl, 2005; Häubl & Trifts, 2000; Payne, Bettman, & Johnson, 1988). This research stream also argues that different types of information (i.e., context variables) and different types of information structures (i.e., task variables) require different levels of effort to process, often measured by the "time per acquisition" (Bettman et al., 1993; Ha & Hoch, 1989; Lurie, 2004; Lynch & Ariely, 2000; Shugan,

1980).

1.2.2. Individual Differences—Demographic Characteristic

Kassarjian (1981) also stated that it is undeniable that there are differences between individuals who, regardless of the product or situation, make some people more interested, concerned or involved in the consumer decision process. Large individual differences in external search intensity have been found to be related to demographic characteristics (Newman, 1977).

1.2.2.1. Education Level

Shim and Drake (1990) reported that, regardless of product category, online shoppers tend to be characterized as having higher educational levels. Previous research has found education to be related positively to search behavior (Claxton, Fry, & Portis, 1974; Newman & Staelin, 1972; Westbrook & Fornell, 1979). According to Wikipedia, an encyclopedia, higher education is the education provided by universities, vocational universities and other collegial institutions that award degree. Consumers who receive higher education usually have more chance to know about the Internet and use the Internet to search for information. According to Eastman and Lyer (2004), consumers with higher levels of education are willing to use the Internet and make online purchases. Education increases the buyer's ability to use information wisely and therefore his/ her need for information. Or as Westbrook and Fornell (1979) have stated: Education was assumed to increase the buyer's need for information related to the purchase decision and thereby to increase the value of search and the likelihood of reliance on high value, high

cost sources such as Consumer Reports and related buying guides, as well as extensive visits to retail outlets.

1.2.2.2. Age

Ratchford, Talukdar & Lee (2001) reported that online purchasers were generally younger; more educated and had higher incomes. Dholakia and Uusitalo (2002) found that younger consumers reported more hedonic (for fun) and utilitarian (with a goal in mind) benefits of online shopping than older consumers. All these five researchers did not study online information search or online purchase behavior but studied the benefits of online shopping only. To fill this gap, in our study we include age as an explanatory variable of search effort and willingness to pay.

1.2.2.3. Income

Research also indicated relationships between search effort and social class, deteriorating economic condition, age, income and mobility (Bucklin, 1969; Katona & Mueller, 1955; Newman & Staelin, 1972; Kiel, 1977). It is difficult to separate the influence of income from general socioeconomic status. Kassarjian (1981) has implied a positive relationship between socioeconomic status and purchasing involvement, and in fact describes his "low involvement" consumer as being a member of the lower socioeconomic class. This would lead to the assumption that higher income might be associated with higher purchasing involvement. The positive relationship found between income and search effort (Claxton, Fry & Portis, 1974; Newman & Staelin, 1972) would provide some indirect support for this notion. However, it would seem that the marginal utility of purchasing involvement

would be low for high income groups; this is because they can purchase almost anything they want and value their free time more than the money that they could save by wise purchasing. Thus it seems that a curvilinear relationship could be expected between purchasing involvement and income, with moderate levels of income producing the highest levels of purchasing involvement and low and high-income groups being relatively less involved.

1.2.2.4. Family structure

Relationship between family structure (i.e., presence of children) and purchasing involvement is apparent. The presence of children is expected to lead to the greatest purchasing involvement. The purchasing involvement conceptualized here is expressed by the search activity. This is true partially because discretionary income is low in these stages (Wells & Gubar, 1966) and the act of purchasing becomes more personally relevant since wise (value oriented) buying is necessary to achieve the family's expected standard of living. At this point it should be noted that education, income, and stage of family life cycle are all related.

Furthermore, multivariate analysis reveals that income, education, age and family structure are important social determinants of online access and that Internet use is the lowest among single mothers, members of lower socioeconomic groups (Bucy, 2000). Pastore (2001) claimed, “Initially the Web audience was populated by the young, affluent and well educated.”

Overall, we can see the various demographic characteristics impact the search effort of the consumers. Previous studies have also looked at demographics as predictors, and purchase intention as the outcome variable (Kim et al., 2004; Lin &Yu, 2006; Kwak et al., 2001).

1.2.3. Hedonic and Utilitarian Attributes

As Dhar and Wertenbroch (2000) documented, consumer choices are driven by hedonic and utilitarian considerations. When consumers are facing a choice of new cell phones, they may care about hedonic features like color and shape, or they may also care about utilitarian attributes such as battery life and sound volume. Research suggests that these different considerations of attributes of products can affect consumers' evaluation and attitudes, and can also enable people to distinguish between goods according to their hedonic or utilitarian nature (Batra & Ahtola, 1990; Mano & Oliver, 1993). To be consistent with the previous research, we use the term hedonic to refer to the aesthetic, experiential, and fun benefits (Dhar & Wertenbroch, 2000; Strahilevitz & Myers, 1998), and the term utilitarian to refer to the functional, instrumental and practical benefits (Hirschman & Holbrook, 1982). In our study we include movies and console games as hedonic goods and tools and health products as utilitarian goods.

Although the consumption of many goods involves the varying of both dimensions (Batra & Ahtola, 1990), usually consumers consider some products as primarily hedonic while others as primarily utilitarian. Hedonic and utilitarian goods can be differentiated on various dimensions

----------Insert Table 1 about here----------

Hedonic goods are really hard for the consumers to judge the quality of the goods prior to the purchase (Sawhney & Eliashberg, 1996). Search of information will not significantly reduce the uncertainty. In comparison, utilitarian goods may be judged on the basis of objective attributes (for the consumption of cellphone, we can easily judge the battery life, price). Thus, utilitarian goods depend on functional and objective attributes, which can be easily evaluated using the information while hedonic goods possess more intangible, symbolic attributes, which are harder to compare even with information (Addis & Holbrook, 2001; Kahnx et al., 1997).

Also there is high consumption risk associated with hedonic products resulting from the uncertainty of quality, subjective attributes and social risks consumers face due to the high emotional involvement and symbolic value behind the hedonic goods (Miller & McIntyre, 1993). In contrast, utilitarian goods bear very low social risks. The purchase motive of hedonic goods depends on variety, emotions and symbolic characters while utilitarian goods can justify the choice on the basis of objective product features (Kahnx et al., 1997).

Hedonism and utilitarianism can also be constructed as a similar but different pair of constructs: wants and shoulds (Bazeman, Tenbrunsel & Wade, 1998). The wants are affectively appealing than the shoulds. So it is more difficult to justify spending on hedonic goods than the utilitarian goods (Prelec & Lowewenstein, 1998; Thaler 1980). Hedonic goods offer benefits in the form of experiential enjoyment while the utilitarian

goods offer benefits in practical functionality (Batra & Ahtola, 1990; Hirschman & Holbrook, 1982; Mano & Oliver, 1993).

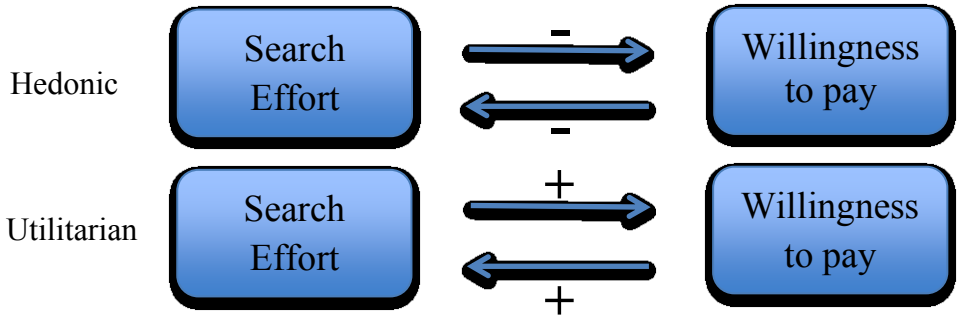
Information on hedonic products will naturally tend to focus more on the fun aspects of the products. For example, a search for console games over the Internet will highlight many ads that focus on the enjoyable experience, the fantastic visual effects and so on. Past research clearly indicates that hedonic consumption evokes a sense of guilt (Khan, Dhar & Wertenbroch, 2004; Kivetz & Simonson, 2002; Prelec & Loewenstein, 1998; Strahilevitz & Mayers, 1998). It is logical to expect that as consumers search more on the Internet, the more they will get to know about the fun that the hedonic products can bring to them; this is likely to make them feel even guiltier. Previous research indicates that the guiltier individuals feel, the less they are willing to pay (Yi & Muhn, 2013). While for utilitarian goods, the more searches the consumers perform, the more solid information consumers can use to construct reasons for justifications of the consumptions (Shafir, Simonson & Tversky, 1993); it is logical to expect that they will thus spend after they thoroughly search the webpage information. Similar logic for the impact of purchase behavior on search behavior happens here. For hedonic goods, the more consumers are considering to spend, the less they are likely search to avoid the information to trigger the sense of guilty. While again for the utilitarian goods, the more consumers are considering to spend, the more information they want to gain to help them justify the purchase decision.

Then we come up our second hypothesis, which are also the main points of focus of this paper.

H2: The product category moderates the interdependence of the search effort and willingness to pay.

H2a: For the hedonic products, there is a negative relationship between search effort and willingness to pay.

H2b: For the utilitarian products, there is a positive relationship between search effort and willingness to pay



1.3. Modeling Framework

Our objective is to propose a model to estimate the interdependence between search effort and willingness to pay. Interdependence in our context is defined as the direct impact of search effort on consumers' willingness to pay, and the direct impact of consumers' willingness to pay on search effort. Under this situation, the explanatory variables are jointly determined with the dependent variable, typically through an equilibrium mechanism. The best way to model this is using a system of two simultaneous equations.

Based on the theories and common knowledge, we define that search effort as duration/page and willingness to pay as (actual total spending for product category)/quantity. The impact between these two factors is direct and happening simultaneously. They are both endogenous variables because of the correlation with the error terms. All the demographic variables are used as the instrument variables for these two endogenous variables.

Consider a model where we observe search effort of the consumers (S_{ij}), and willingness to pay (P_{ij}) of product j on personal i. We can model the interdependence between these two variables by the following simultaneous equations:

$$S_{ij} = X_{ij}*\beta_1 + P_{ij}*Z_1 + \varepsilon_1$$

$$P_{ij} = Y_{ij}*\beta_2 + S_{ij}*Z_2 + \varepsilon_2$$

The explanatory variable that is determined simultaneously with the dependent variable is generally correlated with the error term, which leads to bias and inconsistency in OLS. We consider the two-equation structural model and focus on estimating the first equation. To show that P_{ij} is generally correlated with ε_1, we solve the two equations for P_{ij} in terms of exogenous variables and the error term. If we plug the right-hand side of the first equation in for S_{ij} in the second equation, we can get:

$$P_{ij} = Y_{ij}*\beta_2 + (X_{ij}*\beta_1 + P_{ij}*Z_1 + \varepsilon_1)*Z_2 + \varepsilon_2$$

Or we can put it in this way:

$(1-Z_1Z_2)* P_{ij} = Y_{ij}*\beta_2 + X_{ij}*\beta_1 Z_2 + \varepsilon_1 * Z_2 + \varepsilon_2$

Then we can rewrite above equation as:

$P_{ij} = Y_{ij}*_2 + X_{ij}* + v_2$

In this equation, P_{ij} is expressed in terms of the exogenous variables and the error terms. This is the reduced form equation for P_{ij}. The reduced error term v_2 is a linear function of the structure error terms, ε_1 and ε_2. Since ε_1 and ε_2 are each uncorrelated with X_{ij} and Y_{ij}, v_2 is also uncorrelated with X_{ij} and Y_{ij}. Therefore, we can consistently estimate by OLS.

A reduced form for S_{ij} also exists and the algebra is similar.

From the reduced form equation, we can tell that v_2 is a linear function of ε_1 and ε_2, so it is correlated with ε_1. Then we can say P_{ij} and ε_1 are correlated because of simultaneity, so OLS estimation of simultaneous equations will produce biased and inconsistent estimators of β_1 and β_2.

The leading method for estimating simultaneous equation models is the method of using instrumental variables (book chapter). Therefore, we proposed using three-stage-least-square method to estimate the models. The three stage least square estimations can give a better and more efficient estimation than the two-stage-least-square estimation methods proved by Peter Schmidt (1977). It has been proved that the correlations between the error terms and the endogenous variables lead to bias and inconsistency in using OLS estimations and this violates the assumption of OLS that every explanatory variable is uncorrelated with the error term. However we can identify the instrumental variables to

consistently estimate the parameters in the simultaneous equations. The instrumental variables are uncorrelated with the error terms but can explain the endogenous variables very well to replace them in the estimation of the simultaneous equations.

These two equations constitute a Simultaneous Equation Model (SEM). For these types of equations, there are several important features. First, given X_{ij}, Y_{ij}, $\varepsilon_{1,}$ $\varepsilon_{2,}$ these two equations determine $S_{ij,}$ $P_{ij.}$ For this reason, S_{ij} and P_{ij} are the endogenous variables. X_{ij} and Y_{ij} are vectors of explanatory variables that are specific to the search effort and willingness to pay and they are both uncorrelated with the two error terms. The usual identification condition, that there is at least one variable in each of other vector, holds. The Z_1 measures the direct effect of search effort on willingness to pay and the Z_2 measures the direct effect of willingness to pay on search effort. ε_1 and ε_2 are two correlated error terms assumed to follow a bivariate normal distribution; that is $[\varepsilon_1, \varepsilon_2] \sim BVN(0,\sum)$.

The key parameters in the model are z and β. Specifically, the z parameters capture the interdependence between search effort and willingness to pay. β captures the effect of the demographics on both dependent variables.

Vectors X_{ij}, and Y_{ij} contain two types of predictors for S_{ij} and Y_{ij} respectively. Most of the predictors are common in both vectors but each vector includes at least one variable that is not in the other, as is necessary for identification of the two equations. In addition to the demographic information like education, family size, age, income, children presence, the data set in this study also include the Internet connection.

1.4. Data and Empirical Analysis

1.4.1. Data

The dataset used in this study was collected from comScore 2004 disaggregate dataset which captures detailed browsing and buying behavior for 50,000 Internet users across the United States. A device installed in each household with permission from the consumers records the consumer behavior of online buying and searing activities. The dataset is a random sample from a massive cross-section of more than 2 million global consumers who have given comScore explicit permission to confidentially capture their Web-wide browsing and transaction behavior. This panelist-level data is gathered by comScore Networks using a proprietary data collection methodology that enables comScore to passively observe the full details of panelists' Internet activity, including every Web site visited and item purchased. Panelists include purchasers and non-purchasers who were active online during each month of the 2004 calendar year. The unique panelist identifier in this dataset is Machine ID. All demographic information is based upon the associated household. All sessions are aggregated by machine ID in the household, so that individual breakdowns are not available and a particular individual could use more than one machine.

We propose several different summary statistics to examine potential associations across sites. In order to show the comparisons between the hedonic and utilitarian categories more clearly, we only included the purely hedonic category, which includes movies and gaming consoles; the purely utilitarian category, which includes health goods and tools.

We aggregated the purchase variables like product quantity and total cost by site session id. For search behavior variables like duration and pages viewed, I just averaged them by the site session id to get the site session level data to match with our purchase behavior variables.

Overall, we consider eleven variables: search effort, willingness to pay, connection, age, education, income, children, household size, racial, census and origins. Table 2 and Table 3 describe the variables.

----------Insert Table 2 and Table 3 about here----------

As mentioned earlier, the model specification makes controlling for the potential endogeneity of the two dependent variables S_{ij} and P_{ij} necessary. The instrumental variables are all exogenous variables in X_{ij} and Y_{ij} relevant to them plus other demographic variables inside the dataset. Specifically, the instrumental variables I_{ij} contain connection, age, education, income, children, household size, racial, census and origins.

1.4.2. Empirical Analysis

Our analysis shows the significant asymmetric interdependence between the two major variables Search Effort (S_{ij}) and Willingness to Pay (P_{ij}), which also vary statistically across different product types. In addition, some of the demographic characteristics are also significant factors on the two dependent variables. However these are not the focus of the study.

Table 4 and Table 5 summarize the estimation results for our model applied on the four product categories.

Model fit: Over all, the model fits well. The system weighted R^2 for all four product categories are 0.0377 (movie), 0.0302 (game consoles), 0.0389 (health) and 0.0594 (tools) separately. The low R^2 value is quite typical for this type of regression estimation.

Significant interdependence: Across the four product categories, the impact of search effort on willingness to pay and the impact of willingness to pay on the search effort are all statistically significant. This is consistent with our first hypothesis that there is significant interdependence between the online consumers search effort and buying behavior.

Asymmetric interdependence: The estimates in table 3 also show the impact of search effort S_{ij} on consumers' willingness to pay P_{ij} is significantly less than the impact of P_{ij} on S_{ij}. In other words, the dependence of willingness to pay on the search effort is less than the dependence of the search effort on the willingness to pay. This provides evidence that how much money consumers are going to spend can impact their search effort more than the reverse situation.

Differences in interdependence across product categories: When we look at the detailed value of each parameter, we can see there is a negative relationship between search effort S_{ij} and willingness to pay for the two hedonic products movies and console games; while the relationship of willingness to pay and the search effort is positive for

the two utilitarian products health goods and tools. This is consistent with our second hypothesis.

Since the interdependence is the study focus we will not discuss the coefficients of the demographics. They worked well as the instrument variables.

----------Insert Table 4 and Table 5 about here----------

1.5. Discussions and Future Research

There has been lots of research on different aspect of hedonic and utilitarian goods. However examining the effect of the nature of the good on the interdependence of willingness to pay and search effort has never been done. In this article, we present a simultaneous equation model incorporating the major research focus of search effort and willingness to pay together with the demographics of the household. Our theoretical framework includes the impact of search behavior and purchase behavior in the online domain as well as the impact of product categories between hedonic and utilitarian goods.

We derive two sets of hypotheses from our theory. Specifically, the first hypothesis is regarding the interdependence between search effort and willingness to pay. From the empirical study, we find significant evidence to support this hypothesis, indicating that under the online environment, the consumers' online search effort would impact their purchase behavior. More importantly, the willingness to pay will impact how consumers are going to search on the websites. The empirical results also supported our hypothesis

that when people start a purchase session online, they mentally initiate the searching and purchasing process simultaneously.

The second group of hypotheses is regarding the moderating effect of the product types between hedonic and utilitarian goods. We find significant evidence to support our expectations that for the hedonic products, consumers express a negative relationship between the search effort and the willingness to pay while for the utilitarian product categories, the relationship becomes positive.

The research findings in this article could have significant implications for decision makers in website designing associated with the nature of their products. First, because our findings suggest that for hedonic products, there is a negative interdependence between search effort and willingness to pay, the website should be designed with less text content which is time consuming to read but more creative pictures to avoid the guilty feelings. Second, our findings also suggest that for utilitarian goods, there is positive interdependence between search effort and willingness to pay. The website should be designed to provide enough information for the consumers to read about and stay longer on the webpage.

While providing support to our theoretical framework, our results are subject to limitations, which also suggest opportunities for further research. Because we have panel data for a year so we didn't examine the dynamic aspects of the interdependence between search effort and willingness to pay. It would be interesting to further explore how

consumers learn from the previous experience and change their buying and searching behavior accordingly.

1.6. References

Addis, M. & Holbrook, M. B. (2001). “On the conceptual link between mass customization and experiential consumption: An explosion of subjectivity”. Journal of Consumer Behavior, 1(1):50-60.

Andreasen, A. R. (1968). “Attitudes and customer behavior: a decision model, in Perspectives in Consumer Behavior”. Kassarjian, H. H. and Robertson, T.S. (eds), Illinois: Scott, Foreman and Co

Batra, R. & Ahtola, O. T. (1990). “Measuring the hedonic and utilitarian sources of consumer attitudes”. Marketing Letters, 2(2):159-170.

Bazerman, M. H., Tenbrunsel, A. E. & Wade-Benzoni, K. (1998). “Negotiating with yourself and losing: Understanding and managing competing international preference”. Academy of Management Review, 23 (2):225-241.

Bettman, J. R., Johnson, E. J., Luce, M. F. & Payne, J. W. (1993). “Correlation, conflict, and choice.” Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(4):931-951.

Bucklin, L. P. (1969). "Consumer search, role enactment, and market efficiency". Journal of Business, 42(4):416-438.

Bucy, E. P. (2000). “Social access to the Internet”. Harvard International Journal of Press/Politics, 5(1):50-61.

Chaney, I. M. (2000). “External Search Effort for Wine”. International Journal of Wine Marketing, 12(2):5 – 21.

Claxton, J. D., Fry, J. N. & Portis, B. (1974). “A taxonomy of prepurchase information gathering patterns”. Journal of Consumer Research, 1(3):35-42.

Dhar, R., & Wertenbroch, K. (2000). “Consumer choice between hedonic and utilitarian goods”. Journal of Marketing Research, 37(1):60-71.

Dholakia, R. R. & Uusitalo, O. (2002). “Switching to electronic stores: consumer characteristics and the perception of shopping benefits”. International Journal of Retail & Distribution Management, 30(10):459-469.

Diehl, K. (2005). “When two rights make a wrong: Searching too much in ordered environments*”.* Journal of Marketing Research, 42(3):313-322.

Eastman, J. K. & Lyer, R. (2004). “The elderly's uses and attitudes towards the Internet”. Journal of Consumer Marketing, 21(3):208-220.

Ha, Y. W. & Hoch, S. J. (1989). “Ambiguity, processing strategy, and advertising-evidence interactions”. Journal of Consumer Research, 16(3):354-360.

Hammond, K., McWilliam, G. & Diaz, A. N. (1998). “Fun and work on the web: differences in attitudes between novices and experienced users”. Advances in Consumer Research, 25(1):372-378.

Häubl, G. & Trifts, V. (2000). “Consumer decision making in online shopping environments: The effects of interactive decision aids”. Marketing Science, 19(1):4-21.

Hirschman, E. C. & Holbrook, M. B. (1982). “Hedonic consumption: emerging concepts, methods and propositions”. The Journal of Marketing, 46(3):92-101.

Johnson, E. J., Bellman, S. & Lohse, G. (2003). “Cognitive lock-in and the power law of practice”. Journal of Marketing, 67(2):62-75.

Kahn, B. E., Ratner, R. K. & Kahneman, D. (1997). “Patterns of hedonic consumption over time”. Marketing Letters, 8(1):85-96.

Kassarjian, H. H. (1981). "Low involvement—a second look". Advances in Consumer Research, 8(1):31-34.

Katona, George & Mueller (1955). "A study of purchase decisions". Consumer Behavior: The Dynamics of Consumer Reaction, Vol. 1, Lincoln H. Clark (ed.) New York: New York University Press, 30-87.

Khan, U., Dhar, R. & Wertenbroch, K. (2005). “A behavioral decision theory perspective on hedonic and utilitarian choice”. Inside consumption: Frontiers of research on consumer motives, goals, and desires:144-165.

Kiel, G. C. (1977). "An empirical analysis of new car buyers' external search behaviour," Ph.D. Dissertation, University of New South Wales.

Kim, E. Y. & Kim, Y. K. (2004). “Predicting online purchase intention for clothing products”. European Journal of Marketing, 38(7):883-897.

Kivetz, R. & Simonson, I. (2002). “Self-Control for the Righteous: Toward a theory of precommitment to indulgence”. Journal of Consumer Research, 29(2):199-217.

Kwak, H., Fox, R. J. & Zinkhan, G. M. (2001). “Factors influencing consumers’ Internet purchases: attitudes, Internet experiences, demographics and personality traits”. American Marketing Association. Conference Proceedings, 12:106-107

Lin, C. H. & Yu, S. F. (2006). “Consumer adoption of the Internet as a Channel: The influence of driving and inhibition factor”. Journal of American Academy of Business, 9(2):112-117.

Lurie, N. H. (2004). “Decision Making in Information-Rich Environments: The Role of Information Structure”. Journal of Consumer Research, 30(4):473-486.

Lynch Jr., J. G. & Ariely, D. (2000). “Wine online: Search costs affect competition on price, quality, and distribution”. Marketing Science, 19(1):83-103.

Mano, H. & Oliver, R. L. (1993). “Assessing the dimensionality and structure of the consumption experience: Evaluation, feeling, and satisfaction”, Journal of Consumer Research, 20(3):451-466.

Miller, C. M., McIntyre, S. H. & Mantrala, M. K. (1993). “Toward formalizing fashion theory”. Journal of Marketing Research, 30(2):142-157.

Newman, J. W. (1977). "Consumer External Search: Amount and Determinants". Consumer and Industrial Buying Behavior, Woodside, A. G., Sheth, J. N. & Bennett, P. D., (ed.). New York: North Holland.

Newman, J. W. & Staelin, R. (1972). “Prepurchase information seeking for new cars and major household appliances”. Journal of Marketing Research, 9(3):249-257.

Okada, E. M. (2005). “Justification effects on consumer choice of hedonic and utilitarian goods”. Journal of Marketing Research, 42(1):43-53.

Pastore, M. (2001). “Women maintain lead in Internet use”. Retrieved July 18, 2002.

Payne, J. W., Bettman, J. R. & Johnson, E. J. (1988). “Adaptive strategy selection in decision making”. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3):534-552.

Peterson, R. A. & Merino, M. C. (2003). “Consumer information search behavior and the Internet”. Psychology & Marketing, 20(2):99-121.

Prelec, D. & Loewenstein, G. (1998). “The red and the black: Mental accounting of savings and debt”. Marketing Science, 17(1):4-28.

Ratchford, B. T., Talukdar, D. & Lee, M. S. (2001). “A model of consumer choice of the Internet as an information source”. International Journal of Electronic Commerce, 5(3):7-21.

Sawhney, M. S. & Eliashberg, J. (1996). “A parsimonious model for forecasting gross box-office revenues of motion pictures”. Marketing Science, 15(2), 113-131.

Schmidt, P. (1977). “Estimation of seemingly unrelated regressions with unequal numbers of observations”. Journal of Econometrics, 5(3):365-377.

Shafir, E., Simonson, I. & Tversky, A. (1993). “Reason-based choice”. Cognition, 49(1):11-36.

Sherif, M. & Cantril, H. (1947). The Psychology of Ego Involvement, New York: Wiley.

Shim, S. & Drake, M. F. (2006). “Consumer intention to utilize electronic shopping. The Fishbein Behavioral Intention Model”. Journal of Direct Marketing, 4(3):22-23.

Shugan, S. M. (1980), “The Cost of Thinking”. Journal of Consumer Research, 7(2):99-111.

Strahilevitz, M. & Myers, J. (1998). “Donations to charity as purchase incentives: How well they work may depend on what you are trying to sell”. Journal of consumer research, 24(4):434-446.

Thaler, R. (1980). “Toward a positive theory of consumer choice”. Journal of Economic Behavior & Organization, 1(1):39-60.

Wells, W. D. & Gubar, G. (1966). “Life cycle concept in marketing research”. Journal of Marketing Research, 3(4):355-363.

Westbrook, R. A. & Fornell, C. (1979). “Patterns of information source usage among durable goods buyers”. Journal of Marketing Research, 16(3):303-312.

Yi, Y. J. (2013). “Mixed Products: How Adding Different Attributes Influences Consumer Perceptions and Product Evaluation”. Asia Marketing Journal, 15(1):83-105.

Table 1 Hedonic Versus Utilitarian Goods

Variable	Hedonic Goods[a]	Utilitarian Goods[b]
Quality uncertainty	Relatively high	Relatively low
Attributes	Subjective, symbolic, intangible	Objective, Functional, tangible
Consumer Risk	Relatively high	Relatively low
Purchase motives	Emotional, variety seeking,	Rational, practical functionality
Type of purchase	Wants	Shoulds

[a]For example, movies, console games. [b]For example, health products, tools.

Table 2 Demographics

Most Educated Head of the Household

0	Less than a high school diploma
1	High School diploma or equivalent
2	Some college but no degree
3	Associate degree
4	Bachelor's degree
5	Graduate degree
99	Missing

Household Income

1	Less than 15k
2	15k-24.999k
3	25k-34.999k
4	35k-49.999k
5	50k-74.999k
6	75k-99.999k
7	100k+

Age of Eldest Head of Household and Age of User

1	18-20
2	21-24
3	25-29
4	30-34
5	35-39
6	40-44
7	45-49
8	50-54
9	55-59
10	60-64
11	65 and over

Household size

1	1
2	2
3	3
4	4
5	5
6	6+

Racial Background

1	white
2	Black
3	Asian
5	other

Connection Speed

0	Not broadband
1	Broadband

Census Region of Residence

1	north east
2	north central
3	south
4	west

Country of origin

0	Hispanic
1	Not Hispanic

Child presence

0	no
1	yes

Table 3 Definitions of Data Variables

Variable	Operationalization
Endogenous Variables	
Search Effort Sij	Duration of web-view divided by pages viewed
Willingness to Pay Pij	Money spending divided by quantity purchased
Exogenous Variables (Predictors Xij, Yij)	
Education	Most educated person in the household are recoded into 2 groups (before college, college and higher)
Family size	Size of the family are regrouped into 2 groups (1-3 small family; 4 and over big family)
Age	Eldest head of household are regrouped into three groups (young family below 30, middle aged family between 30 and 54; old family above 54)
Income	Household income are regrouped into three groups (low income family below 24.999k, medium income family between 25k to 49.999k, high income family above 75k)
Children presence	Two groups (with children or not)
Connection	Two groups (broadband or not)
Exogenous Variables (Instrumental Variables I_{ij})	
Racial	4 groups (White, Black, Asian, other)
Origin	2 groups (Hispanic or not)
Census	4 groups (north east, north central, south, west)

Table 4 Three-stage least-square estimation results of willingness to pay as DV

		Movie	Console Games	Health	Tools
System weighted R^2		0.0377	0.0302	0.0389	0.0594
Endogenous variable					
S_{ij}		-0.166[a]	-0.775[a]	0.333[a]	0.665[a]
Education	Ed1	-0.016	-0.091[a]	0.0179	-0.074
	Ed2	0.018	-0.093[a]	0.0461[a]	0.045
Family size		0.007	0.115[a]	0.013	0.031
Age	Age1	-0.033[a]	0.021	0.029	-0.105[a]
	Age2	-0.005	0.045	-0.001	-0.072
Income	In1	-0.030	0.028	-0.082[a]	-0.116
	In2	-0.016	-0.096	-0.040	-0.085
Children		-0.009	0.02	-0.030	0.109[a]

[a] significant at 0.05 level

Table 5 Three-stage least-square estimation results of Search effort as DV

		Movie	Consoles Games	Health	Tools
System weighted R^2		0.0377	0.0302	0.0389	0.0594
Endogenous variable					
P_{ij}		-1.082[a]	-0.457[a]	1.454[a]	1.11[a]
Education	Ed1	-0.012	-0.084[a]	-0.011	0.123[a]
	Ed2	-0.026	-0.086[a]	-0.061[a]	-0.051
Family size		-0.008	0.082[a]	-0.03	-0.04
Age	Age1	-0.061[a]	-0.01	-0.073[a]	0.127[a]
	Age2	-0.027	0.017	-0.044	0.091
Income	In1	0.051[a]	0.085	0.166[a]	0.116
	In2	0.048[a]	-0.029	0.096[a]	0.072
Children		-0.011	0.008	0.059[a]	-0.142[a]
Connection		-0.181[a]	-0.063[a]	-0.125[a]	-0.048[a]

[a] significant at 0.05 level

Chapter/Essay 2

Analyzing Online User Comments: A Text Mining Approach

2.1. Introduction

It is widely known that user recommendations play a role in affecting potential purchasers of products and services. This is one reason why user reviews are prominent in numerous web sites (e.g., at Amazon.com). Kumar and Benbasat (2006) for example, use empirical evidence to demonstrate the influence of recommendations and online reviews on consumers' perceptions of the usefulness and social presence of such websites. Reichheld (2003) in fact essentially argues that recommendations are the single biggest predictors of company growth. Senecal and Nantel (2004) found that individuals who consulted online product recommendations selected recommended products twice as often as individuals who did not consult recommendations. On the whole it seems quite clear that recommendations should be of very high value for marketers.

Since recommendations are so crucial for firms, it is also important for marketers to identify what might be some of the drivers of recommendations. There have been many studies that have tried to identify variables that affect the decisions of individuals to make recommendations of product or not. Some examples of such research are Lowenstein (1995), Brown et al. (2005), and Shabbir et al. (2007). However, almost all such previous research has used potential determinants' data that was numeric in nature. As a consequence only a limited number of determinants could be studied in any particular

study (e.g., Ladhari et al. 2011 identified three drivers of recommendations– perceived service quality, emotional satisfaction and image).

There is an abundance of consumer reviews in the online domain. Among these plentiful online reviews there are also recommendations made by consumers. The reviews themselves are typically in text form. This also means that such reviews can contain a large number of potential determinants of recommendations. The wide availability of lengthy and numerous text-based online reviews provides a treasure trove of information that can potentially reveal a much wider set of variables that determine whether a recommendation is made or not. This is certainly a largely unexamined issue in the marketing literature. In order to extract such rich information, there is a need to use text-mining models. This is a goal of our study--to investigate in detail how a study of such online reviews can reveal determinants of user recommendations.

In the marketing literature, researchers have studied online reviews. Several research papers have looked at the impacts of reviews on variables such as sales and preferences using only numerical data as independent variables. For example Chevalier and Mayzlin (2006) have looked at volume (number of ratings) and valence (average numerical ratings) of reviews across online book retailers to see their effect on sales. The valence of reviews is an important concept and the authors have used numerical data related to the product (books in this case) to capture valence. Dellarocas et al. (2007) in the marketing literature also found a similar relationship between volume of ratings and box office revenues. Almost all such papers in the traditional marketing literature include analysis of

numerical information such as those derived from rating scale responses, and almost no work on text mining models exists in the marketing literature.

Potential consumers search online for information and for evaluating choice alternatives. These individuals have access to a large number of sentence based product reviews from previous consumers. The data content of these reviews is very different from traditional structured numerical data; an entirely different modeling approach therefore needs to be utilized to extract relevant information from these reviews.

With the help of text mining, which mainly handles unstructured data/text, we can actually investigate online content more deeply than done previously in the marketing literature. Machine-learning algorithms can be used for categorizing text material (Apte & Damerau, 1994; Lewis et al., 1996; Dagan et al., 1997; Sebastiani, 1999). These algorithms can be employed to classify texts—the huge heterogeneous, unstructured data available especially on the web --- like reviews and blogs into fixed categories such as that reflecting sentiment polarity based on the content of the text messages.

In the present research we use a machine language algorithm to extract information from text-based online reviews to reflect consumer perceptions that drive user recommendations. The amount of information contained in text-based comments is tremendously greater than that contained in typical numerical rating type data. There are such a large number of types of permutations and combinations of words that are possible. Hence a major challenge is how to extract the most relevant information from

this very big data source. In this paper we attempt to extract the essence of the information present in consumer reviews by using a text mining methodology.

One needs to be aware that extraction of valid information from such online reviews is associated with examining many words and also large numbers of combinations of words. Overall the amount of text that one has to wade through is very high, and therefore, the issue of scalability becomes very important in such an investigation. In our research, we investigated scalability by doing the following steps. First we compared the time needed for data preprocessing with how it is related to the number of reviews considered and the number of words evaluated. We do a similar analysis with respect to the time needed for classification also. We found that the time of computation is not at all large even for the very large dataset that we used; this indicates good scalability for our methodology. Additionally we compare running our models using a serial algorithm (which uses a single core on a processor) vs. using a multi-threading parallel algorithm (which can use multiple cores on a processor to shorten the processing time). We find that utilizations of the parallel algorithm reduce computation time tremendously; this means that our methodology will become even more scalable.

In our research, we extracted a large amount of data from the Internet; this extraction itself is very computing-intensive, and we wrote specific Java programs to do this. After we extracted the data from the web site, we wrote and used another series of Java programs to conduct pre-processing, indexing, feature selections etc. in order to get the data formatted for doing classification analysis. The algorithms were implemented in Eclipse of version Juno Service Release 2. Eclipse is an open-source Integrated

Development Environment (IDE) for Java. The dependent variable is whether users recommended the product or not while the independent variables in our predictive model were different combinations of words from online reviews. We used a specific structured process to identify the most relevant key words and key word combinations out of the many possible ones. Predictive accuracies of this text-based model were high across a variety of examined situations.

A new finding from our empirical modeling is that as the number of textual features increases, the predictive accuracy of the model also increases but only up to a point. Beyond that, inclusion of more features in the model leads to a decrease in predictive accuracy. In essence, we found an inverted-U shape of relationship between the number of features and model accuracy. This finding also has positive implications from the perspective of scalability in the analysis of big data of this kind.

Text based data has the potential to provide valuable diagnostic information about what individuals are thinking; focus on developing such diagnostics is not common in the marketing literature. We develop an approach, which lends structure to such diagnostics. This is based upon identifying groups of common and unique word features. We used two different types of processes to identify which word combinations are most important. One process relies on the weighted frequency of words. The other approach identifies determinant words by computing their discriminating ability using Chi-square value calculations. This second approach provides very good insights with respect to understanding whether consumers make recommendations or not; this kind of approach has not been used previously in the business literature for such diagnostic purposes. We

empirically compare the nature of this diagnostic information across product types (hotel types in our case) and across consumer segments (business or personal trip). Identification of such keyword based diagnostic information is of substantial value for search engine marketing like Google AdWords.

Identification of attributes affecting recommendations via usage of text mining models together with a detailed investigation of the importance of the *scalability* of this text-based problem analysis, we believe, has not been studied in the marketing literature. We also believe that there is a tremendous potential for usage of the methodology of this study for many future areas of study in the academic marketing discipline.

The rest of the paper is organized as follows: in section 2.2 we discuss the fundamentals of our text modeling methodology. In section 2.3 we discuss the details how we implement this for online reviews and recommendations. Section 2.4 provides a summary of our findings and suggests future directions for research in marketing.

2.2. Methodology

In this section we describe the overall approach that we use for analysis of text content. In section 3 we give more specifics of how this approach was used with our online database of text.

Text classification will use a machine-learning algorithm to classify the sentence based text documents into one of previous defined categories. Suppose we have a set of documents which could be the reviews posted on the websites by consumers. Each

document can be expressed as a vector of attributes$X = (X_1, X_2, \cdots, X_n)$. All documents belong to one of several predefined categories$Y = (Y_1, Y_2, \ldots, Y_m)$. The attributes are usually term weights from indexing which will be discussed in detail in the following sub sections. We will use a model to predict the class of the document$Y = f(X)$. In our research, we focus on two classes: recommend or not recommend.

In the following sections, we will explain the complete process of how we used text mining.

2.2.1. Preprocessing

Before a learning method can be applied, a number of preprocessing steps are required to get the data in ready format for further analysis. The preprocessing of raw data includes: raw text tokenization, case conversion, stop-words removal and stemming.

First, the raw text is divided into tokens (single word, special symbols, etc.) using whitespaces (space, tab, new line character, etc.) as separators to break the entire review document into tokens. For example, suppose we have a document stating *"I like iPhone. It is the first phone I got and I really like the appearance."* The tokenization step will break this sentence into tokens like *"I", "like", "iPhone", "got"* etc.

The second step is case conversion where the words are modified to be all in lower cases—all the capitalized letters will be converted into lower cases. In the above example, the letter *"P"* is converted to *"p"* and the word *"iPhone"* is converted to *"iphone"*. The

purpose of case conversion is to reduce the number of redundant words by converting them all into lower cases.

The third step is removal of stop-words. The purpose of stop-words removal is to reduce the size of the classification matrix by reducing the number of irrelevant terms. Lots of very commonly used words like *"the", "I", "to"*, etc., are of little use in classifying documents into predefined categories. The efficiency and accuracy of the classifications can be improved by removing these words. In our study a general stop-word list, which contains standard stop words, was used along with some manual adaptations.

The next preprocessing step is called stemming. Different variations of a word are converted into a single common form that is termed stem. For example, "*connect"* is the stem for "*connected", "connection", "connecting"*, etc. Usage of stemming significantly reduces the number of features and increases retrieval performance (Kraaij & Pohlmann, 1996). We use a dictionary-based stemmer to do stemming with our data. When a term is unrecognizable, we use logic to give the word a correct stem.

2.2.2. Indexing

The result so far is a term-by-document matrix with each cell representing the raw frequencies of occurrence for each term in each document. The rows of the matrix represent terms (words), and the columns represent documents (reviews for example). Jones (1972) showed that there is a significant improvement in retrieval performance by using weighted terms vectors. The term weight is generated by multiplying Term

Frequency (TF) and the Inverse Document Frequency (IDF) (Jones, 1973; Coussement & Van Den Poel, 2008b).

TF measures the frequency of the occurrence of an indexed term in the document (Salton & Buckley, 1988; Coussement & Van Den Poel, 2008a). The higher the frequency of a term, the more important this term is in characterizing the document. Such frequency of occurrence of an indexed word is used to indicate term importance for content representation (Baxendale, 1958; Luhn, 1957; Salton & McGill, 1983).

In our study, the TF was obtained from the raw term frequency. Not every word appears equally across the whole set of review documents. Some words appear more frequently than others by nature. Given other things constant, the more seldom a term occurs in a document collection, the more distinguishing strength that term is likely to have. Hence the weight of a term is inversely proportional to the number of documents in which it appears (Coussement & Van Den Poel, 2008a). So IDF is used to take into account of this effect. The logarithm of the IDF will decrease the effect of the raw IDF-factor (Coussement & Van Den Poel, 2008a).

Finally the total weight of a term i in document j is given by $w_{ij} = TF_{ij} \times IDF_i$

Here, TF_{ij} is equal to the term frequency of term i in document$j$; IDF_i is equal to the inverse document frequency of term$i$.

Mathematically, $TF_{ij} = n_{ij}$ with n_{ij} being equal to the frequency of term i in document j and $IDF_i = log_2\left(\frac{n}{df_i}\right) + 1$, with n being equal to the total number of documents in the

entire collection of reviews and df_i equals to the number of review documents where term i was present (Coussement, 2008).

2.2.3. Multi-word Phrases

Tokenization gives the term-by-document matrix. Each term in the matrix is a single word. In most cases, multi-word phrases are also important because phrases have more complete context information than the individual word. So the most popular class of features used for text classification is n-grams (Pang et al., 2002; Wiebe et al., 2004). Word n-grams include the single word (unigram), and higher order n-grams like bi-grams and tri-grams. Word n-grams have been used effectively in various studies (Pang et al., 2002). Unigram to tri-grams have typically been used in text mining and large n-gram phrase sets require the use of attribute selection to reduce the dimensionalities (Abbasi et al., 2008; Ng et al., 2006). For instance, if we have a sentence *"I like iPhone"*. We have three unigrams *"I", "like", "iPhone"*, two bi-grams *"I like", "like iPhone",* and one tri-gram *"I like iPhone"*.

2.2.4. Dimensionality Reduction

So far this weighted term-by-document matrix is a high dimensional matrix since there are many unique terms. Moreover, it is very sparse with many zeros since not all documents contain all terms (Coussement, 2008). It is worth noting that large attribute dimensionality incurs high computational costs and can cause over-fitting problems in the classification process. So we need to reduce the dimensionality. The number of terms can be reduced through feature selection, which selects a subset of the top-ranked features based on various algorithms. Information-theoretic measures such as chi-squares,

information gain and gain ratios are commonly used in text classification (Sebastiani, 2005) for feature selection. These measures are designed to measure the dependency between the class and the term (Chou et al., 2010). Yang and Pedersen (1997) reported that information gain and chi-squares outperform other functions such as mutual information etc. Debole and Sebastiani (2004) reported that gain ratio and Chi-squares are more effective than information gain. So we choose Chi-squares as the method for attribute selection in our present research.

Chi-square is a common statistical test that measures the lack of independence of two variables (Liu & Setiono, 1995), which are class of document and a feature in the case of text classification. As is well known in traditional statistics, the chi-square test can check the independence of two events A and B. A and B are independent if $P(AB) = P(A) \times P(B)$. For selecting terms/words, the two events are occurrence of the term and the occurrence of the class. In order to get the chi-square values, we need to first build a 2×2 contingency matrix per class-term pair. Suppose we have only two classes: 0 (negative) and 1 (positive). For each term the observed frequency value is:

Observed	Class = 1	Class = 0
Term t appears t = 1	N_{11}	N_{10}
Term t not appear t = 0	N_{01}	N_{00}

Then we need to get the expected frequency value by the equation:

$$E_{11} = N \times P(t) \times P(c) = N \times \frac{N_{11} + N_{10}}{N} \times \frac{N_{11} + N_{01}}{N}$$

Expected	Class = 1	Class = 0
Term t appears t = 1	E_{11}	E_{10}
Term t not appear t = 0	E_{01}	E_{00}

The chi-square value of each term can be obtained by the equation:

$$\chi^2 = \sum_{\substack{c=i \\ t=j}} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

We then rank the terms with respect to the chi-square values. A high value leads to rejection of the independence hypothesis. If the two events are dependent, existence of the term makes the existence of the class more probable. Thus this term will help to discriminate the class of the document.

2.2.5. Classification Technique

We use the Support Vector Machine (SVM) approach for classification purposes. The following discussion draws from Turney and Pantel (2010). SVM was developed by Salton (1971) and Salton, Wong, and Yang (1975). SVM represents each document in a collection as a point in a space. Points in close proximity in the space are grouped into the same categories while points that are far from each other are grouped in to a different class. SVMs are linear classifiers that find a hyperplane to separate two classes of data, positive and negative. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class.

Let X be a term-document matrix. Suppose the document collection contains n documents (in our study, it is the number of reviews) and m unique terms (obtained after

data preprocessing and feature selection). Let w_i be term i in the vocabulary and let d_j be document j in the collection. The row i in X is the row vector $x_{i,:}$, and the column j in X is the column vector$x_{:,j}$. The row vector $x_{i,:}$ contains n elements, one element for each document, and the column vector $x_{:,j}$ contains m elements, one element for each term. The element x_{ij} in X is the TF-IDF of the i-th item w_i in the j-th document. The pattern of numbers in $x_{i,:}$ is a kind of signature of the i-th termw_i; likewise, the pattern of numbers in $x_{:,j}$ is a signature of the j-th document. The notations and descriptions in this paragraph follow from the work by Turney and Pantel (2010).

A simple case (Cristianini & Shawe-Taylor, 2000) of using SVM for classification is shown in Figure 1 for illustrative purposes only:

----------Insert Figure 1 about here----------

Even for the SVM classification method, there are various algorithms and the most popular one is the Sequential Minimal Optimization (SMO), which is conceptually simple, easy to implement and fast to compute. Since computational theory is not the focus of our study and has already been developed by previous researchers, here we only present the idea conceptually; for further details, Cristianini and Shawe-Taylor (2000) is a good reference.

2.2.6. Evaluation Criteria

For assessing the performance of different classification models, we use three criteria: the percentage accuracy, F-measure and the Area Under the receiving operating Curve (AUC) (Coussement and Van Den Poel, 2008a)

Accuracy: This essentially refers to the percentage correctly classified. If *TP*, *FP*, *TN*, *FN* are respectively the number of correctly predicted positive reviews, the number of negative reviews predicted as positive, the number of correctly predicted negative reviews, and the number of positive reviews predicted as negative, accuracy is defined as $(TP + TN)/(TP + FP + TN + FN)$. The accuracy can be compared to the proportional chance criteria ($\text{percentage}_{\text{positive}}^2 + (1\text{-percentage}_{\text{positive}})^2$) in order to confirm the predictive capabilities of a classifier (Morrison, 1969).

F-measure (Powers, 2007): Another way to evaluate the performance of the prediction model is the F-measure. When we look at the performance of the models, we can get the Precision *p*, which is the number of correctly classified positive examples divided by the total number of examples that are classified as positive and the Recall *r*, which is the number of correctly classified positive examples divided by the total number of actual positive examples in the test set separately (Liu, 2008). In order to look at them together, the F-measure is used to combine the precision and recall as the harmonic mean of precision and recall.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F = \frac{2 \times (precision \times recall)}{precision + recall}$$

A higher F-value indicates better model performance.

AUC (Metz, 1978)**:** In order the get the AUC, we need to first draw the Receiver Operating Characteristics (ROC) curve. This considers the sensitivity *TP / (TP+FN)* and 1 minus the specificity *(1-TN / (TN+FP))* in a two-dimensional graph (Coussement and Van Den Poel, 2008a). The sensitivity is the likelihood of identifying a positive case when presented with one while the specificity is the likelihood of identifying a negative case when presented with one (Gopal et al., 2007). The area under this ROC curve is calculated to compare the performance of a binary classifier (Hanley & McNeil, 1982). A classifier can produce a single ROC point. If a learning algorithm produces the classifier, changing the class ratio in the training set can generate a series of ROC points. Then we can connect all the different ROC points in a figure. In turn, the AUC can be calculated. When classifying randomly, the ROC curve is a line joining points (0, 0) and (1, 1) with the area under the curve equals 0.5. In general, any classification performance should be better than a randomly made classification. A general example (Metz, 1978) of the ROC curve is shown in Figure 2 for illustrative purposes only.

----------Insert Figure 2 about here----------

2.3. Empirical Analysis

In order to test if consumer recommendations can be predicted using only text data, we gathered data from the hotel industry and conducted an empirical evaluation.

2.3.1. Data Collection

In our study, we use data obtained from orbitz.com, which is a leading website in the travel industry. We focus on the hotel domain. On the websites, consumers need to register to leave their reviews, ratings, and recommendation choices after they stayed in the hotel. We collect the data using our own Java program of four levels of hotels in Las Vegas: a 2-star hotel "American Best Value Inn", a 3-star hotel "Bally", a 4-star hotel "Treasure Island" and a 5-star hotel "Venetian". For each hotel, we further segment the reviews into business and personal based on the reviewers' choice of trip purpose. We choose Las Vegas among the various cities across the whole nation because it is one of the most popular tourist cities in the U.S., and attracts a large number of hotel users. Table 6, Table 7, Table 8 and Table 9 represent the summary of the review data of the four hotels.

----------Insert Table 6, Table 7, Table 8 and Table 9 about here----------

Reviews from the first available one to all those posted on the website till April 1st, 2013 were collected. For each review, the reviewer gave the choice of recommend or not-recommend; we use this as the dependent variable (class) in the prediction models. The various sets of words/features are used as the predictive variables.

2.3.2. Empirical Analysis

The first part of the analysis is the prediction of the sentiment polarity of the reviews. We use recommend or not-recommend given by the reviewers as our dependent variable and form this binary text classification problem. We perform data preprocessing to get a high dimensional tri-grams term-by-review matrix. We use TF-IDF to do the term weighting. We then perform feature selection to reduce the dimensions. For the feature selections, we use the selection criteria chi-square as stated previously to help us form a series of different sized feature sets. For the purpose of classification we use the SVM algorithm. We use a common 10-folds cross validation for classification testing and prediction. The classification model confirms our expectation that there is information content in the reviews that can help predict consumers' overall attitude toward the hotel: whether the consumers recommend it or not.

The second part of the analysis is a key focus of our analysis: diagnostics of the key features. We used two different types of processes to identify which word combinations are most relevant. One process relies on the weighted frequency of words. In previous research the frequency of occurrence of an indexed word has been used to indicate term importance for content representation (Baxendale, 1958; Luhn, 1957; Salton & McGill, 1983). Here we combine the term frequency with the inverse document frequency to incorporate the nature of the reality of the term and so we use TF-IDF of the term to represent the importance of the features. The other approach identifies determinant words by computing their ability to discriminate between the existence of a recommendation or not, using Chi-square value calculations.

In the last part of the analysis, we did the scalability test. While automated classification techniques are at the core of analyzing sentence-based reviews, computational requirements are a challenge in the analysis of very large data sets with tens or even hundreds of millions of records. To evaluate scalability we compared the computing time for indexing (a very time-intensive preprocessing step) and for the classification task, at different numbers of reviews and features. Three hotel datasets, Bally, Treasure Island and Venetian are analyzed. We also compared indexing computation time when either a parallel algorithm or a serial algorithm was used.

2.3.3. Results and Discussions

In this section, we report the performance results of online reviews' classification models. We also present the diagnostics from the text analysis. The scalability issue is also discussed here.

2.3.3.1. Classification Performance

Table 10, Table 11, Table 12, Table 13, Table 14, Table 15, Table 16, Table 17 and Table 18 report the predictive performances of different numbers of features as input for the three levels of hotels: 3-star Bally, 4-star Treasure Island and 5-star Venetian. We report the feature size along with the overall accuracy, F-measure and ROC for the three levels of hotels, and for each hotel we also report the prediction performance across the four consumer segments: business, couple, family and friend.

With the very large amount of text in the reviews on the website, we get a huge term-by-review matrix for each hotel with over tens of thousands of features/words. It is very

necessary to choose a subset of features to perform the prediction classification. We apply the chi-square feature selection to rank the number of features and for further dimension reduction. We rank the features by chi-square value and select a number of top features. For each set of ranked features, we evaluate several different numbers of top ranked features (words), from as small as 10 to as large as 5,000 to get a comprehensive idea of how feature size may affect predictive performance of the model.

Table 10 shows how predictive accuracy changes as the size of features increase- for the entire dataset and for the different segments separately. We can see that all the predictive accuracies are good and also greater than the benchmark proportional chance criteria (Morrison, 1969). Table 11 shows that F-measure values are very high for the whole dataset as well as for the different segments indicating accurate predictions as well. Table 12 shows that ROC is generally greater than what one would get by random chance alone (0.5). Similar patterns can be found in Table 13, Table 14, Table 15, Table 16, Table 17, and Table 18. Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, and Figure 11 are plots of accuracy percentage vs. the number of features (words). Overall we can see an interesting inverted U-shaped prediction performance. Addition of words increases accuracy but further addition leads to a decline in accuracy. This pattern is generally consistent across the three hotels and across the four consumer segments.

----------Insert Table 10, Table 11, Table 12, Table 13, Table 14,

Table 15, Table 16, Table 17 and Table 18 about here----------

----------Insert Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, and Figure 11 about here----------

2.3.3.2. Diagnostics

The second part of the results focuses on the diagnostic use of the text mining methodology. In the previous section, we empirically show that the text mining models can classify the users' recommendations for the hotels very well. What the consumers put on websites can represent their real thoughts about their experience with the hotels. So additionally what is important for us is to discover and identify those text features, which are really important from the viewpoint of providing diagnostic information to the companies. Next we describe two ways by which we identify the limited number of more relevant text features from the reviews.

The TF-IDF value reflecting the frequency of occurrences of the word features indicates the importance of the features for representation of the content of the reviews. Therefore, we can rank the features based on TF-IDF values to indicate the importance of the features and help us identify the top important features. We believe that importance is only one aspect of the features for being determinant attributes of consumers' recommendations. Another aspect of the determinant attributes is their discriminating ability in terms of identifying yes or no consumer recommendations; we use chi-square values to identify the features that provide a high ability to discriminate. We list a limited number of words that are very highly ranked.

First we compare a 2-star and a 5-star hotel to see the differences in relevant features. Table 19 and Table 20 summarize the comparisons. We can see from Table 19, by looking at the TF-IDF value ranked features, 2-star and 5-star hotels have several important features in common, like “locations”, “room”, “service” etc. There are also important features unique to each hotel type. For the 2-star hotel, “value/price/cheap” were important. This finding has face validity since common sense would also indicate that consumers who choose to stay at a 2-star hotel are looking for cheap hotels. For consumers staying at a 5-star hotel, some additional services like “shop”, “get-away”, “show”, “dining”, “casino” and “luxury” rank high in our generated list. Again this is consistent with common sense that consumers who choose to stay at 5-star hotels are looking for a luxury experience and a high level of service.

When looking at the discriminating ability of the features on the basis of chi-square value, Table 20 shows that there are some common features across 2-star and 5-star hotels and some of these are consistent with what we found by using the frequency based TF-IDF approach; these words include “location”, “room”, and “service”. There are also some other features like “rude”, “carpet”, “furniture” which were less frequently used but were much more effective in their ability to make a difference between recommendations and no recommendations.

----------Insert Table 19 and Table 20 about here----------

Table 21, Table 22, Table 23, Table 24, Table 25 and Table 26 show the commonalities and differences in word importance and discriminating power across segments for the

various quality level hotels. Even for the same level of hotels, across different segments, there might be similar and different features, which are important for the consumers and in discriminating the consumers' attitudes toward the hotels. Table 21 of Bally hotel shows that based on the ranking of TF-IDF value, "location", "room", "service", "value", and "comfort" etc. are important to consumers across the business and personal segments (including couple, family, friend). When it comes to the unique features, "conference" and "internet" stand out for the business segment while additional enjoyable vacation related services like "parking", "show", "shop" become important for the personal visits. When it comes to the discriminating feature based ranking methodology, like in Table 22 there are some common features that appear in both business and personal segments like "rude", "staff". There are also unique features appear in business segments like "bathroom", "wall paper", "value" and etc. and in personal segments like "location", "anniversary", "casino" and etc. Similar patterns can be found in Table 23, Table 24, Table 25 and Table 26.

----------Insert Table 21, Table 22, Table 23, Table 24, Table 25 and Table 26 about here----------

Identifications of these key features will help advertisers choose the right words or combinations of words for advertising and especially for search engine based advertising messages.

2.3.3.3. Scalability Tests

First, on the platform of Intel Core i5 1.7 GHz with 8 GB memory, on OS X 10.8.5 operating system, we examined how the time of the major preprocessing step—building the index of each term in the matrix -- varied with the number of reviews or the number of features. The results are shown in Table 27, Table 28 and Table 29 and Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, Figure 19, Figure 20, Figure 21, Figure 22, and Figure 23. Figure 12, Figure 13 and Table 27 show that for 2000 reviews or 61568 features the computation time of indexing is about 740 seconds. Figure 14, Figure 15 and Table 27 show that for 2000 reviews or 61568 features the computation time for classification is only about 0.71 second. In other words, our preprocessing methodology or classification can be applied to very big data sets without too much penalty in computation time, thus indicating good scalability for our methodology.

Table 30 shows the comparison of indexing computation time of the 4 hotels when we use a serial algorithm (utilizing a single core) vs. a parallel algorithm (utilizing 8 cores). All the experiments are performed on Intel Xeon X5472 3.0 GHz with 16 GB memories. As seen in Table 30, taking advantage of multi-cores can greatly reduce the execution time and makes our approach even more powerful and scalable when dealing with big data.

----------Insert Table 27, Table 28 and Table 29 about here----------

----------Insert Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, Figure 19, Figure 20, Figure 21, Figure 22, and Figure 23 about here----------

----------Insert Table 30 about here----------

2.4. Summary and Conclusions

Online reviews of products and services are present all over the Internet. Potential consumers value these greatly. Marketers can also get valuable information from reading these reviews. These reviews predominantly contain text-based information. In our present research we utilize text-mining methodology to develop models where factors related to words are independent variables and the dependent variable is whether a consumer recommends a hotel or does not. We find that our word-based model can very accurately predict whether a recommendation is made or not. In the marketing literature, online reviews have been analyzed in the past but text based modeling of this kind does not seem to exist in the marketing literature. In addition the impacts of words and word combinations on user recommendation patterns have not been studied in the marketing literature.

One of the interesting new findings from our empirical analysis is that as the number of words increases, the predictive accuracy of the above models initially increases. The accuracy peaks at a certain number of words and then decreases; an inverted U-shaped relationship exists. The implication is that one does not need to utilize larger and larger number of words in our text mining model to get high accuracy of prediction; this is a favorable thing from the scalability perspective when handling big data.

In addition to making predictions of recommendations, marketers would benefit tremendously if they can identify key words from many thousands of reviews; we suggest a framework by which companies can get this important *diagnostic* information. This framework consists of reliance on the importance of words based on frequency of occurrence and a new way to look at how certain words have greater power to discriminate/distinguish between existence and non-existence of recommendations. Words identified by this diagnostic approach will be of use to advertising managers when they plan on designing messages appropriate for search engine advertising as in Google Adwords; a single ad here can use only a small number of words, and the choice of the keywords could become crucial from the viewpoint of revenue generation. Identification of a few key words using a discriminatory power based approach has seen almost no application in the marketing research literature, and as just stated has clear managerial implications for the ever growing field of search engine advertising. For a perspective on the size of this field, and how the methodology in this paper can be potentially useful to industry, one may note that the finance and insurance industry spent $4 billion on AdWords in 2011, and Amazon alone spent an estimated $55.2 million on AdWords advertising in 2011 (Gabbert, 2012).

Our empirical analysis that includes predictive models and diagnostics is applied for multiple subcategories (different star levels) within a product category (Hotels), and for four different consumer segments. The general pattern of results with respect to good predictive accuracy and the inverted U shaped relationship was generally consistent across all these different scenarios. The diagnostic information identifying key words or

word combinations with respect to different hotel categories and consumer segments was of course not always the same as is to be expected logically. The identified key words seemed to follow good logic, and thus lend face validity to our analysis and findings. These words would be good determinants of online recommendations.

As is obvious, text data is very large in size. Scalability of models and methodologies is an issue that absolutely needs to be addressed when one is dealing with big data. We do a detailed analysis of this and show (see Tables 9a-9c, 10 and Figures 6a-6d, 7a-7d, 8a-8d) that our methodology is very scalable with the big data that we analyze.

The potential future directions for this research stream are numerous. The overall methodology designed in this paper is a foundation that can be applied to a variety of marketing situations. In this paper we apply the text mining technique for the hotel industry. In the future, this can be applied to any other industry. In today's digital era consumers freely express their opinions about products and services on many websites. This provides numerous information sources that can help academicians and practitioners in analyzing consumer attitudes. Even for the hotel industry, due to time limitations, we have only explored reviews of some hotels in Las Vegas; future research can explore more locations. Besides, we can extend this methodology to study a tremendous variety of research questions that would benefit from the analysis of text content posted by web users all over the Internet. Advertisers and marketers would be among the prime beneficiaries once they can glean the appropriate information from text based reviews.

2.5. References

Abbasi, A., Chen, H. & Salem, A. (2008). "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums". ACM Transactions on Information Systems (TOIS), 26(3):1-34.

Apté, C., Damerau, F. & Weiss, S. M. (1994). "Automated learning of decision rules for text categorization". ACM Transactions on Information Systems (TOIS), 12(3):233-251.

Baxendale, P. B. (1958). "Machine-made index for technical literature: an experiment". IBM Journal of Research and Development, 2(4):354-361.

Brown, T. J., Barry, T. E., Dacin, P. A. & Gunst, R. F. (2005). "Spreading the word: Investigating antecedents of consumers' positive word-of-mouth intentions and behaviors in a retailing context". Journal of the Academy of Marketing Science, 33(2):123-138.

Chevalier, J. A. & Mayzlin, D. (2006). "The effect of word of mouth on sales: Online book reviews". Journal of Marketing Research, 43(3):345-354.

Chou, C. H., Sinha, A. P. & Zhao, H. (2010). "Commercial Internet filters: Perils and opportunities". Decision Support Systems, 48(4):521-530.

Coussement, K. (2008). "Employing SAS text miner methodology to become a customer genius in customer churn prediction and complaint E-mail management". SAS Global Forum.

Coussement, K. & Van den Poel, D. (2008a). "Improving customer complaint management by automatic email classification using linguistic style features as predictors". Decision Support Systems, 44(4):870-882.

Coussement, K. & Van den Poel, D. (2008b). "Integrating the voice of customers through call center emails into a decision support system for churn prediction". Information & Management, 45(3):164-174.

Cristianini, N. & Shawe-Taylor, J. (2000). "An introduction to support vector machines and other kernel-based learning methods". Cambridge University Press.

Dagan, I., Karov, Y. & Roth, D. (1997). "Mistake-driven learning in text categorization". Conference on Empirical Methods in Natural Language Processing.

Debole, F. & Sebastiani, F. (2004). "Supervised term weighting for automated text categorization". Text Mining and Its Applications, 81-97. Springer Berlin Heidelberg.

Dellarocas, C., Zhang, X. M. & Awad, N. F. (2007). “Exploring the value of online product reviews in forecasting sales: The case of motion pictures”. Journal of Interactive marketing, 21(4):23-45.

Gabbert, E. (2012). “25 Fast facts about AdWords”. http://www.wordstream.com/blog/ws/2012/08/13/google-adwords-facts#.

Gopal, K., Sacchettini, J. C. & Ioerger, T. R. (2007). “Database approaches and data representation in structural bioinformatics”. Bioinformatics and Bioengineering. Proceedings of the 7th IEEE International Conference on IEEE, 425-432.

Hanley, J. A. & McNeil, B. J. (1982). “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. Radiology, 143(1):29-36.

Jones, K. S. (1972). “A statistical interpretation of term specificity and its application in retrieval”. Journal of documentation, 28(1):11-21.

Jones, K. S. (1973). “Index term weighting”. Information Storage and Retrieval, 9(11):619-633.

Kraaij, W. & Pohlmann, R. (1996). “Viewing stemming as recall enhancement”. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 40-48.

Kumar, N. & Benbasat, I. (2006). “Research note: the influence of recommendations and consumer reviews on evaluations of websites”. Information Systems Research, 17(4):425-439.

Ladhari, R., Souiden, N. & Ladhari, I. (2011). “Determinants of loyalty and recommendation: The role of perceived service quality, emotional satisfaction and image”. Journal of Financial Services Marketing, 16(2):111-124.

Lewis, D. D., Schapire, R. E., Callan, J. P. & Papka, R. (1996). “Training algorithms for linear text classifiers”. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 298-306.

Liu, B. (2008). http://www.cse.lehigh.edu/~brian/course/2008/webmining/presentations/classification.pdf

Liu, H. & Setiono, R. (1995). “Chi2: Feature selection and discretization of numeric attributes”. 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. IEEE Computer Society. 388-391.

Lowenstein, M. W. (1995). “Customer retention: An integrated process for keeping your best customers”. Milwaukee, WI: ASQC Quality Press. 105-114.

Luhn, H. P. (1957). “A statistical approach to mechanized encoding and searching of literary information”. IBM Journal of research and development, 1(4):309-317.

Metz, C. E. (1978). “Basic principles of ROC analysis”. Seminars in nuclear medicine. WB Saunders, 8(4):283-298.

Morrison, D. G. (1969). “On the interpretation of discriminant analysis”. Journal of marketing research, 6(2):156-163.

Ng, V., Dasgupta, S. & Arifin, S. M. (2006). “Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews”. Proceedings of the COLING/ACL on Main conference poster sessions, 611-618. Association for Computational Linguistics.

Pang, B., Lee, L. & Vaithyanathan, S. (2002, July). “Thumbs up?: sentiment classification using machine learning techniques”. Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 79-86. Association for Computational Linguistics.

Powers, D. M. (2007). “Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation”. Journal of Machine Learning Technologies 2(1):37–63.

Reichheld, F. F. (2003). “The one number you need to grow”. Harvard Business Review, 81 (12): 46-54.

Salton, G. (1971). “The SMART retrieval system: experiments in automatic text retrieval”. Information Processing and Management, 24(5):513-523.

Salton, G. & Buckley C. (1988). “Term-weighting approaches in automatic text retrieval”. Information Processing and Management, 24(5):513-523.

Salton, G. & McGill, M. J. (1983). “Introduction to modern information retrieval”. McGraw-Hill, New York.

Salton, G., Wong, A. & Yang, C. (1975). “A vector space model for automatic indexing”. Communications of the ACM, 18(22):613-620.

Sebastiani, F. (1999). “Machine learning in automated text categorization”. Journal of the ACM Computing Surveys, 34(1):1-47.

Sebastiani, F. (2005). “Text mining and its applications to intelligence”. CRM and Knowledge Management, WIT Press, Southampton, UK, 109-129.

Senecal, S. & Nantel J. (2004). “The influence of online product recommendations on consumers’ online choices”. Journal of Retailing, 80(2):159–169.

Shabbir, H., Palihawadana, D. & Thwaites, D. (2007). “Determining the antecedent and consequences of donor-perceived relationship quality – a dimensional qualitative research approach”. Psychology and Marketing, 24(3):271-293.

Turney, P. & Pantel, P. (2010). “From frequency to meaning: vector space models of semantics”. Journal of Artificial Intelligence Research, 37:141-188.

Wiebe, J., Wilson, T., Bunescu, R. & Niblack. W. (2004). “Learning subjective language”. Computational Linguistics, 30(3):277-308.

Yang, Y. & Pedersen, J. O. (1997). “A comparative study on feature selection in text categorization”. Proceedings of ICML-97, 14th International Conference on Machine Learning, 412-420.

Figure 1 SVM Classification

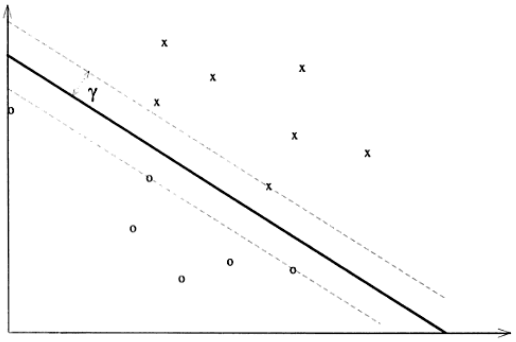


Figure 2 ROC Curves

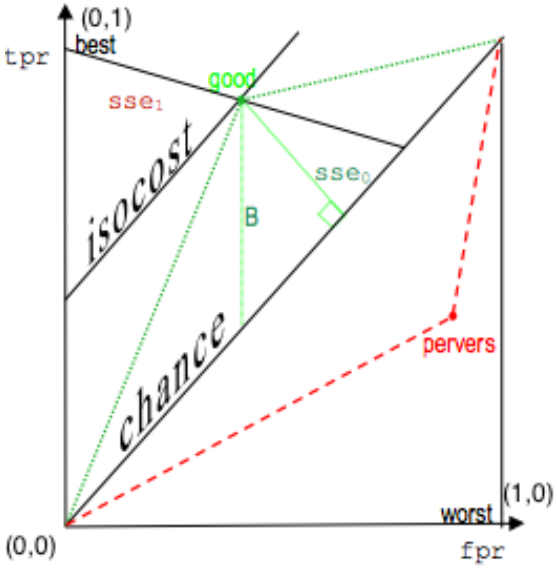


Figure 3 Prediction Results of Bally (Accuracy)

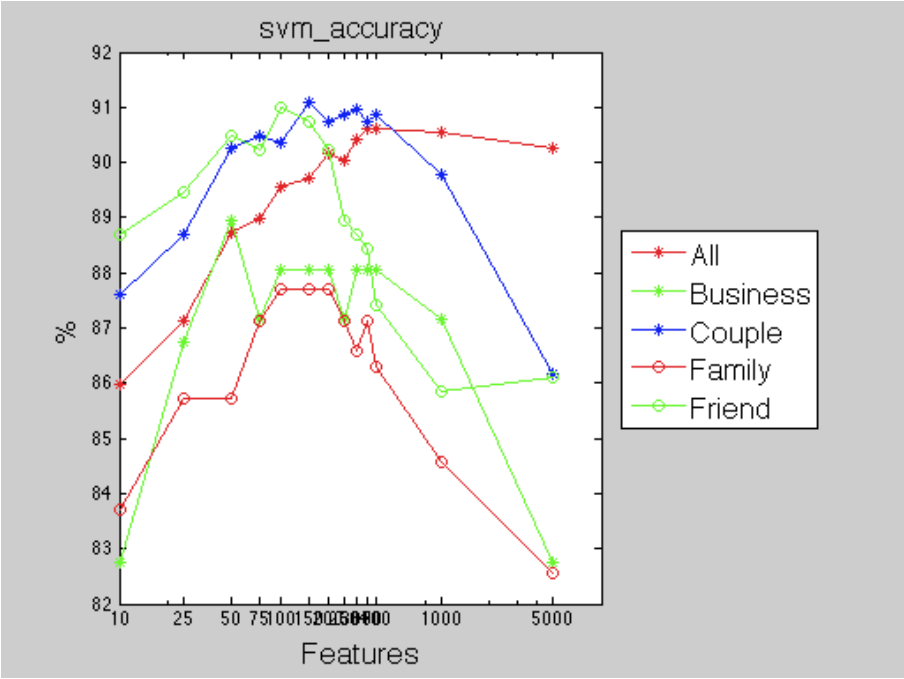


Figure 4 Prediction Results of Bally (F-measure)

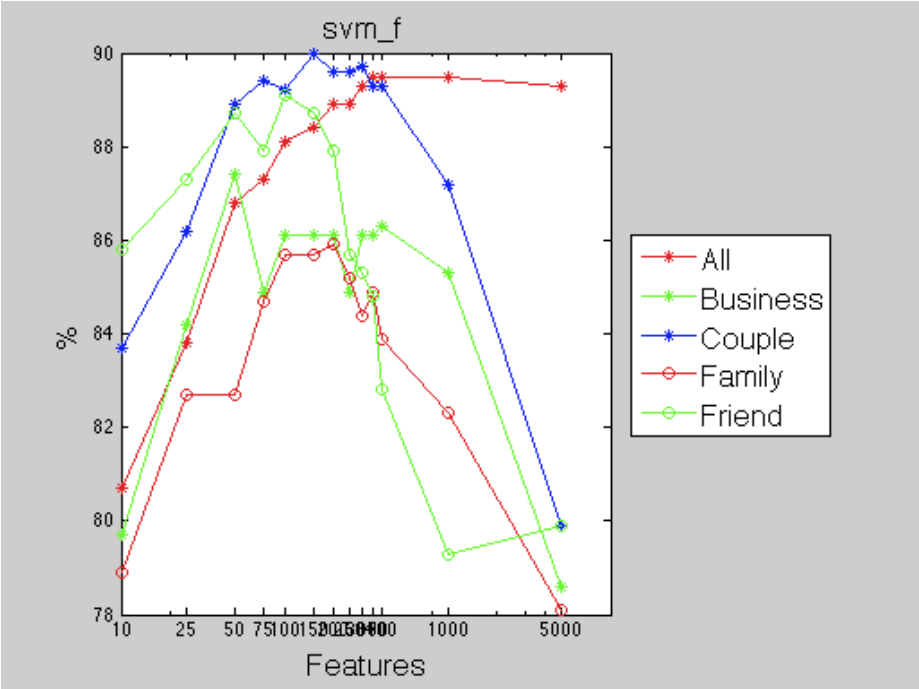


Figure 5 Prediction Results of Bally (ROC)

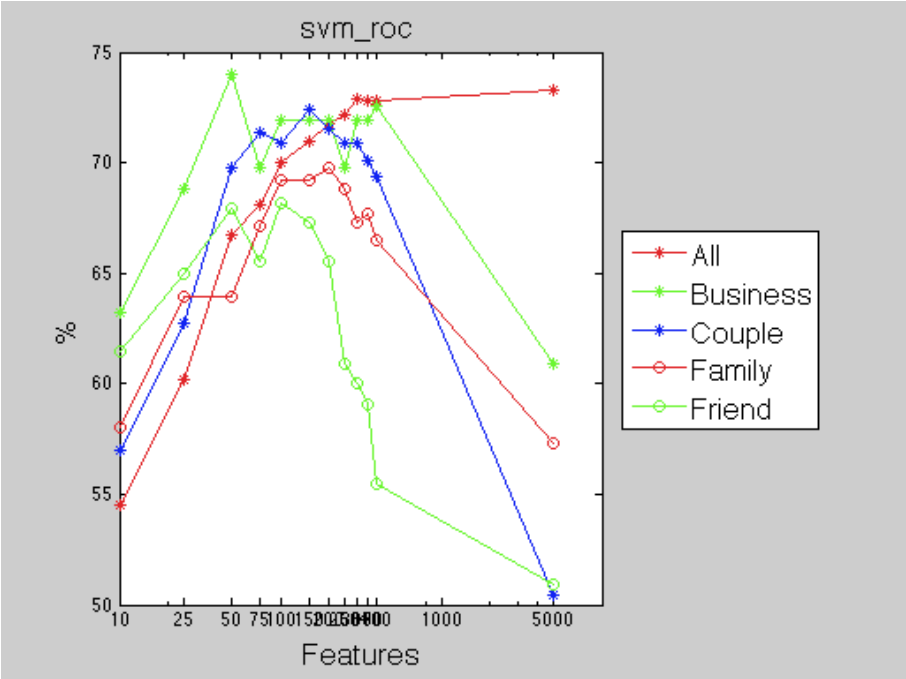


Figure 6 Prediction Results of Treasure Island (Accuracy)

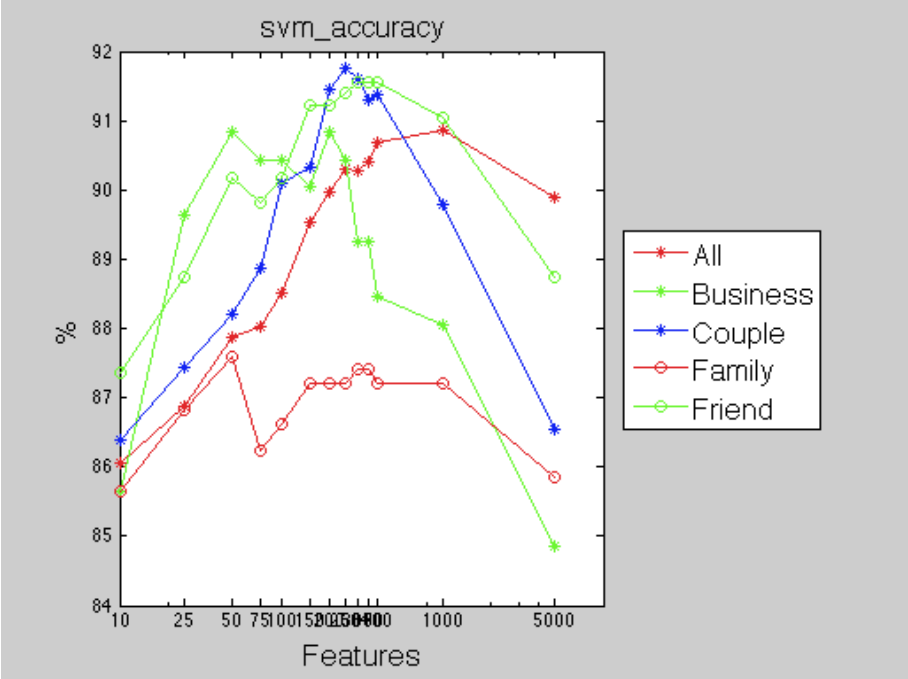


Figure 7 Prediction Results of Treasure Island (F-measure)

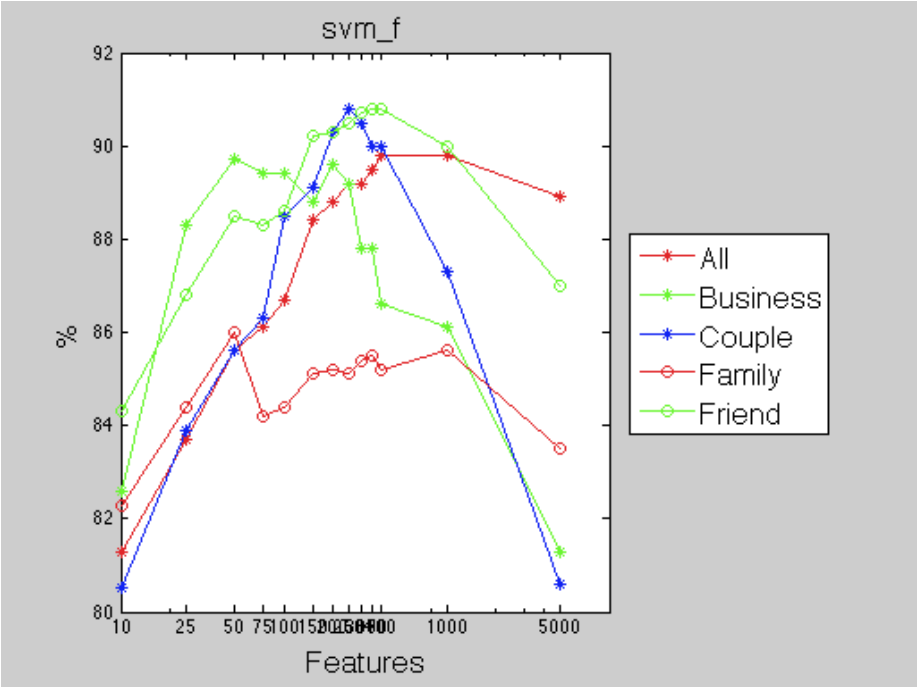


Figure 8 Prediction Results of Treasure Island (ROC)

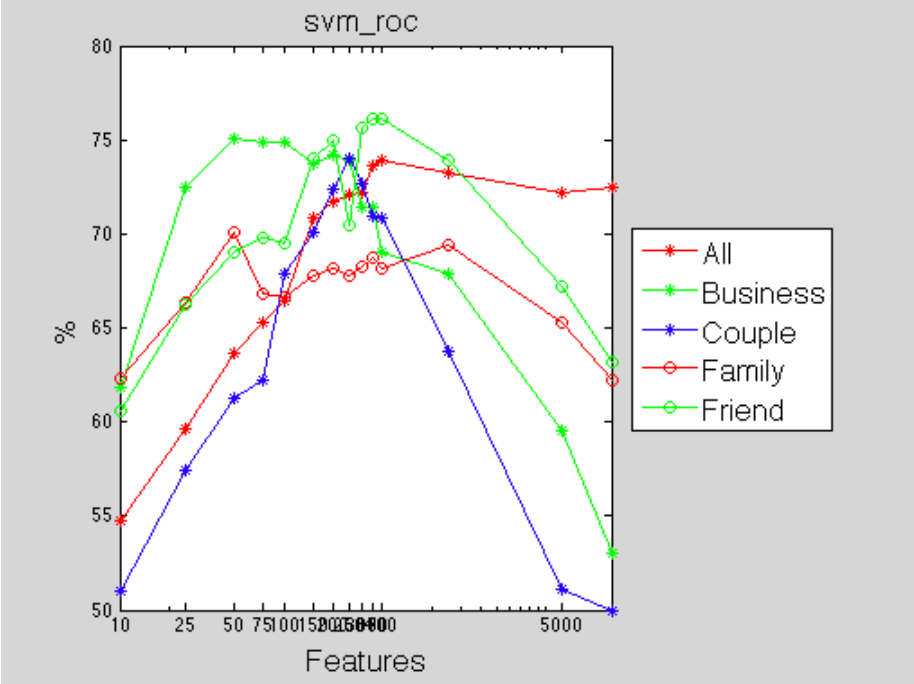


Figure 9 Prediction Results of Venetian (Accuracy)

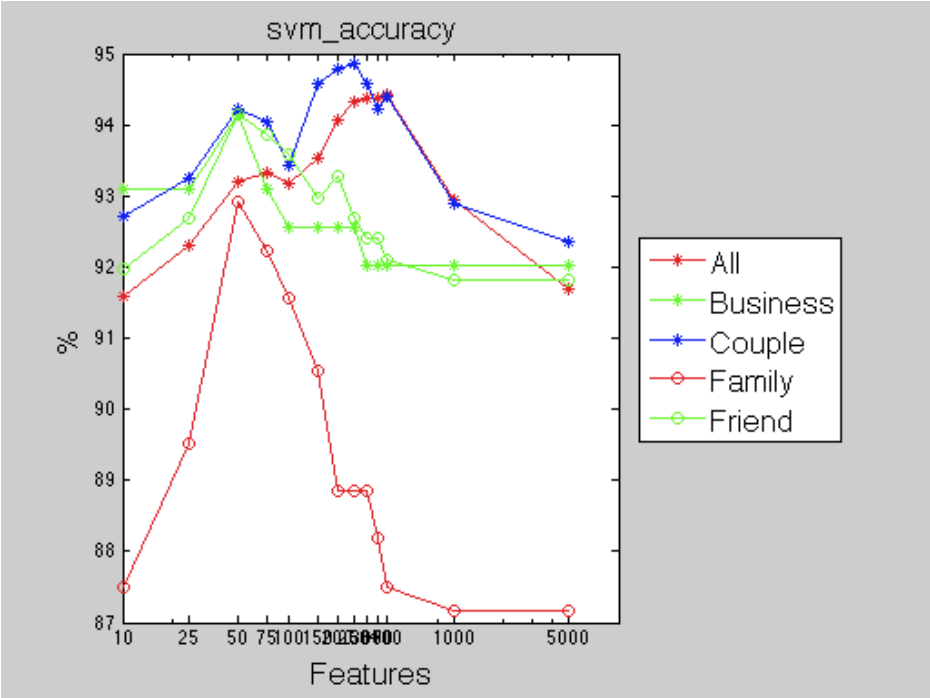


Figure 10 Prediction Results of Venetian (F-measure)

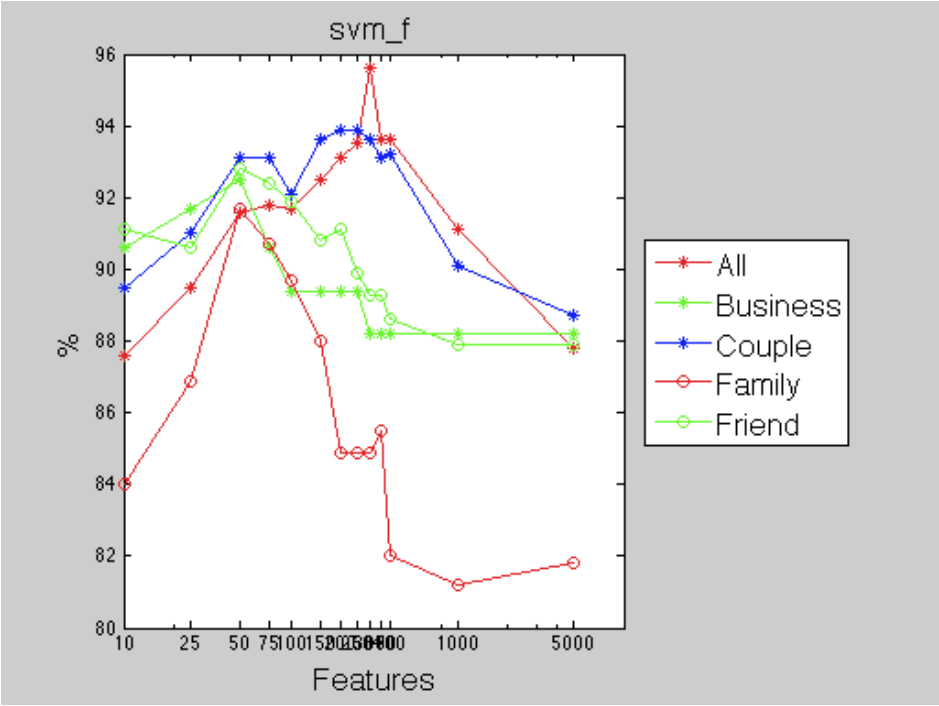


Figure 11 Prediction Results of Venetian (ROC)

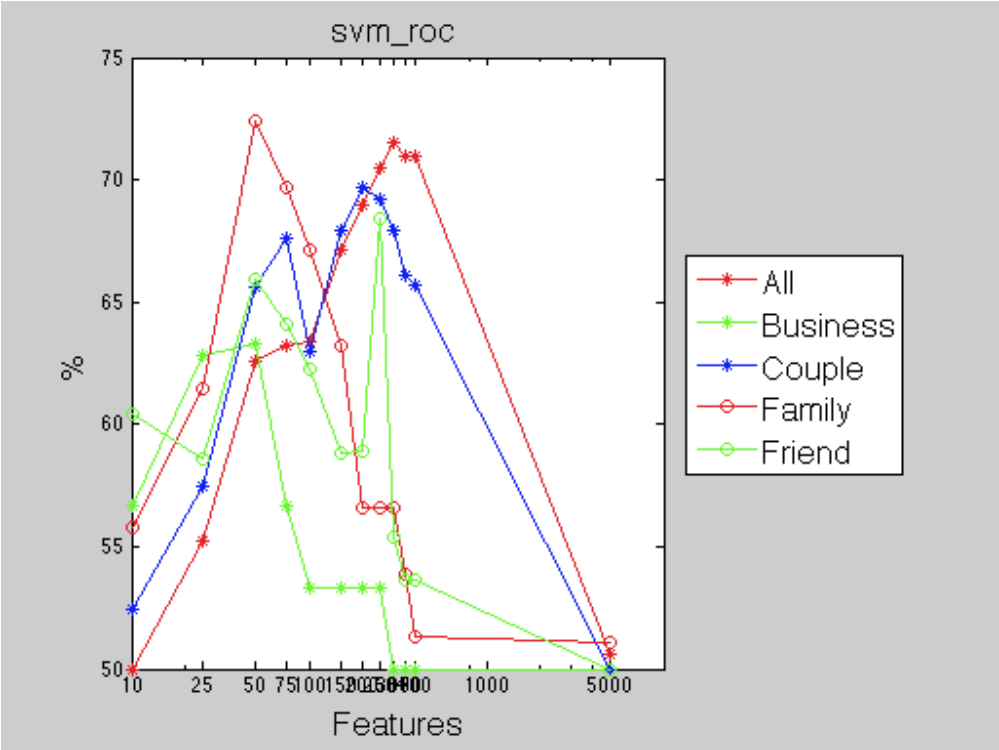


Figure 12 Indexing Computation Time vs. # of Reviews of Bally

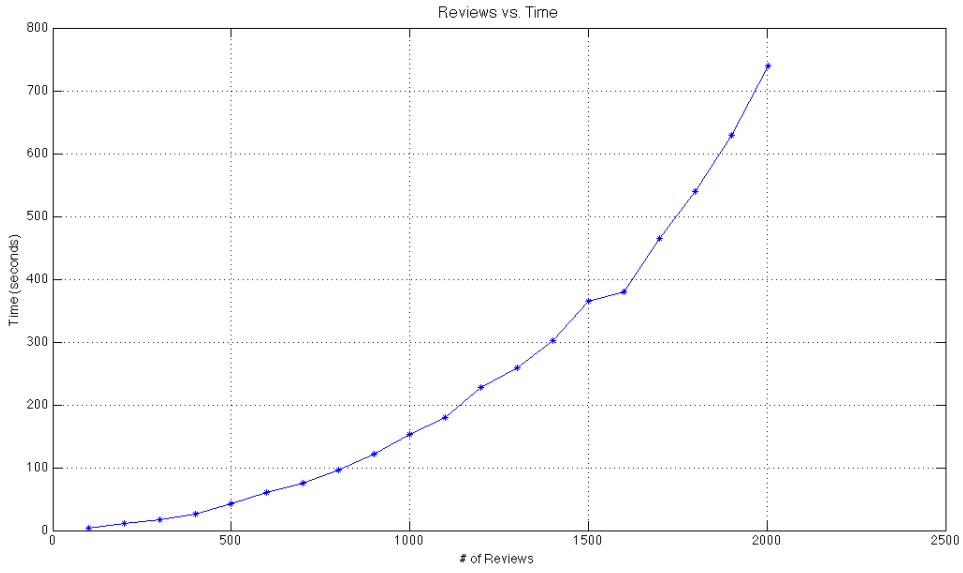


Figure 13 Indexing Computation Time vs. # of Features of Bally

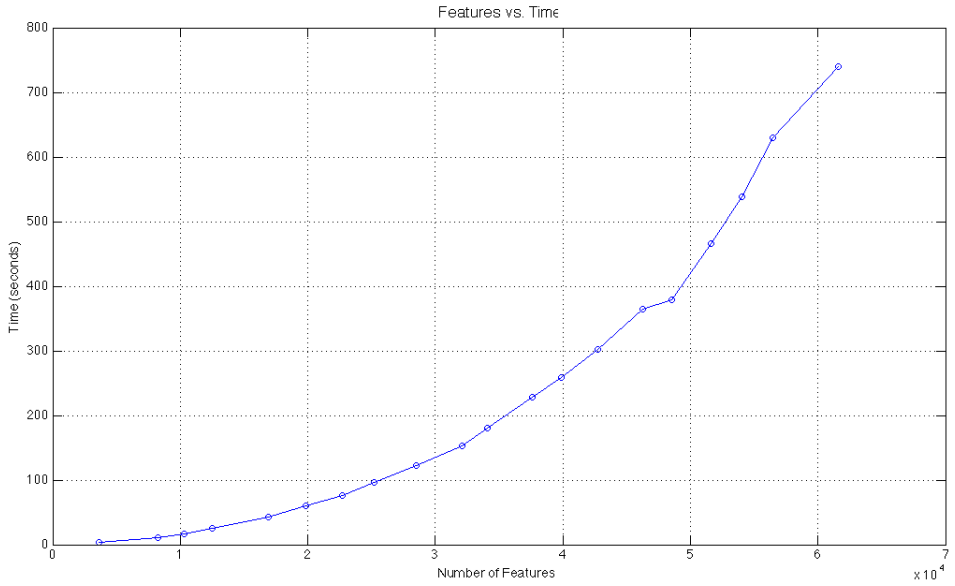


Figure 14 Classification Computation Time vs. # of Reviews of Bally

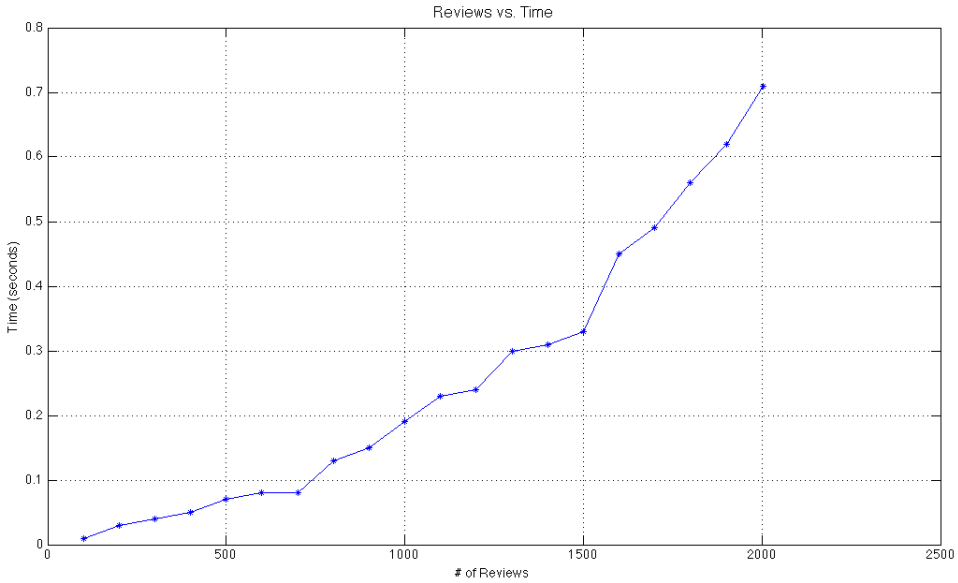


Figure 15 Classification Computation Time vs. # of Features of Bally

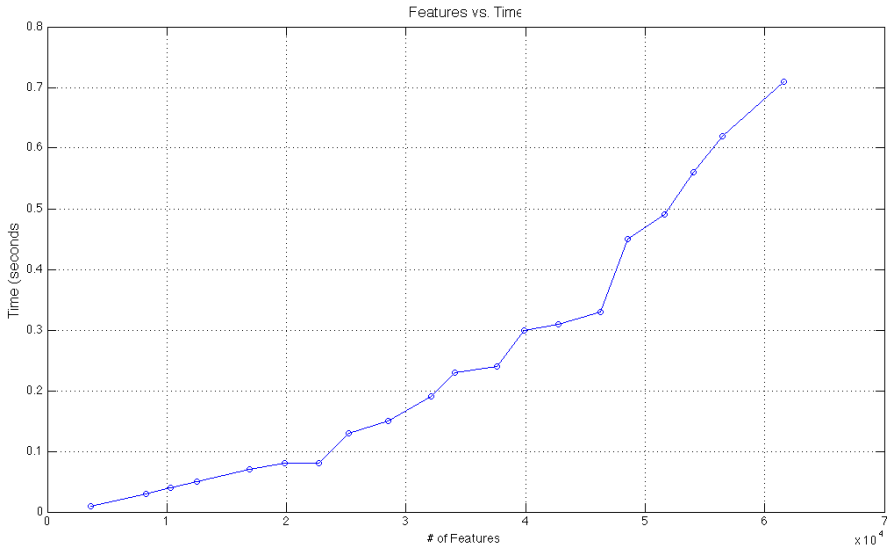


Figure 16 Indexing Computation Time vs. # of Reviews of Treasure Island

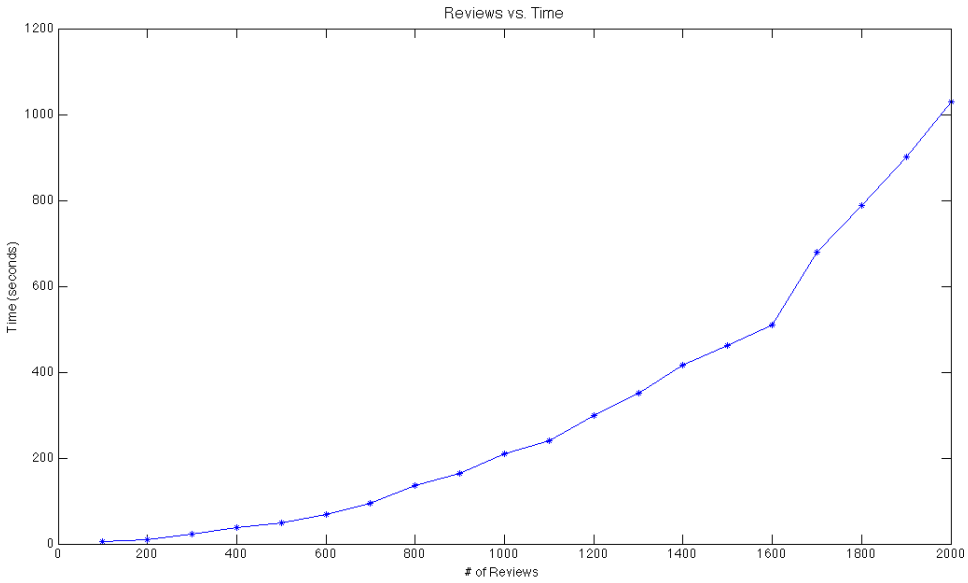


Figure 17 Indexing Computation Time vs. # of Features of Treasure Island

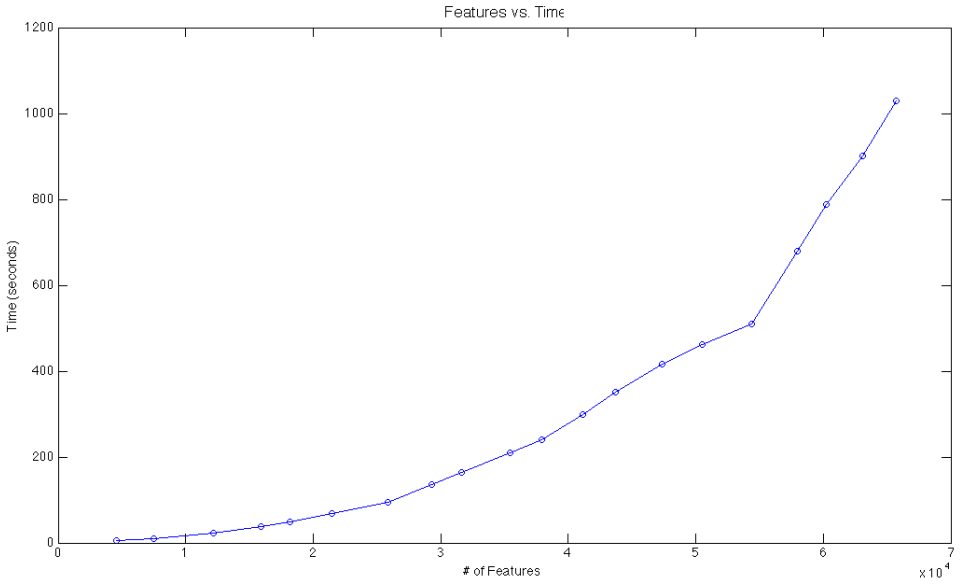


Figure 18 Classification Computation Time vs. # of Reviews of Treasure Island

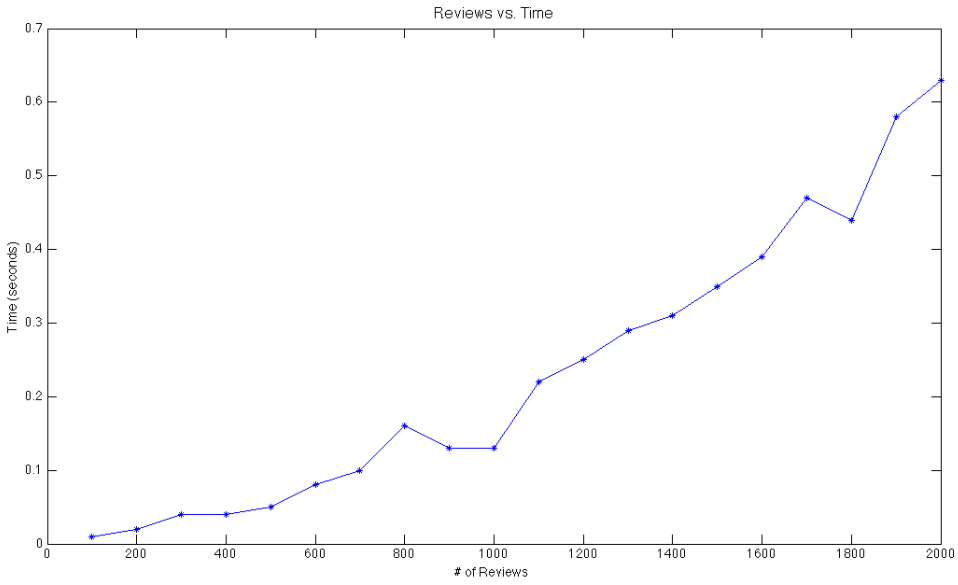


Figure 19 Classification Computation Time vs. # of Features of Treasure Island

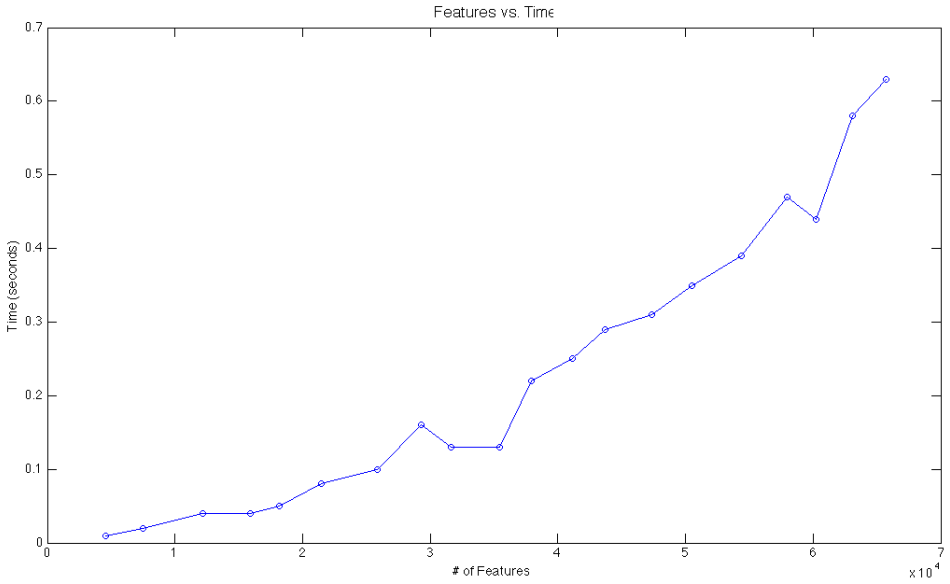


Figure 20 Indexing Computation Time vs. # of Reviews of Venetian

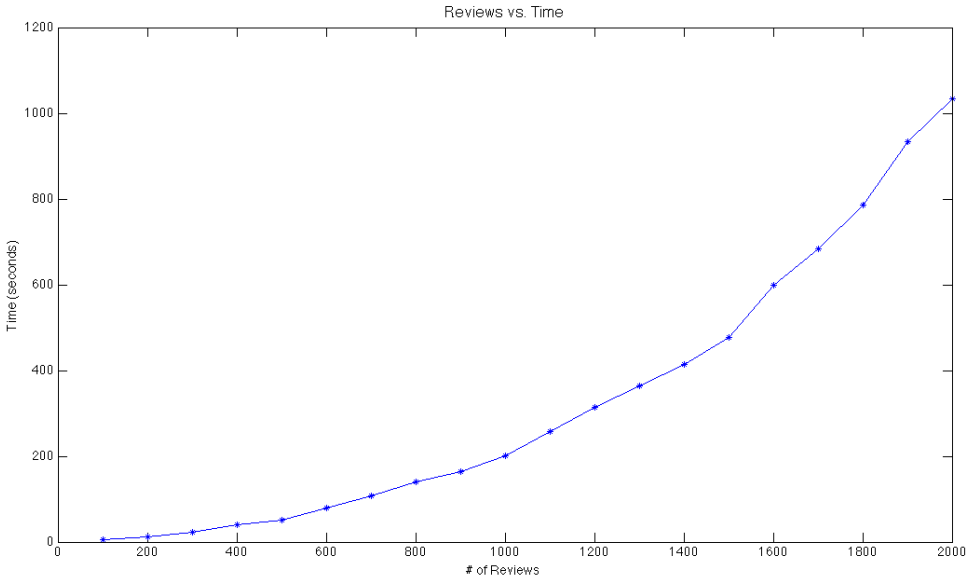


Figure 21 Indexing Computation Time vs. # of Features of Venetian

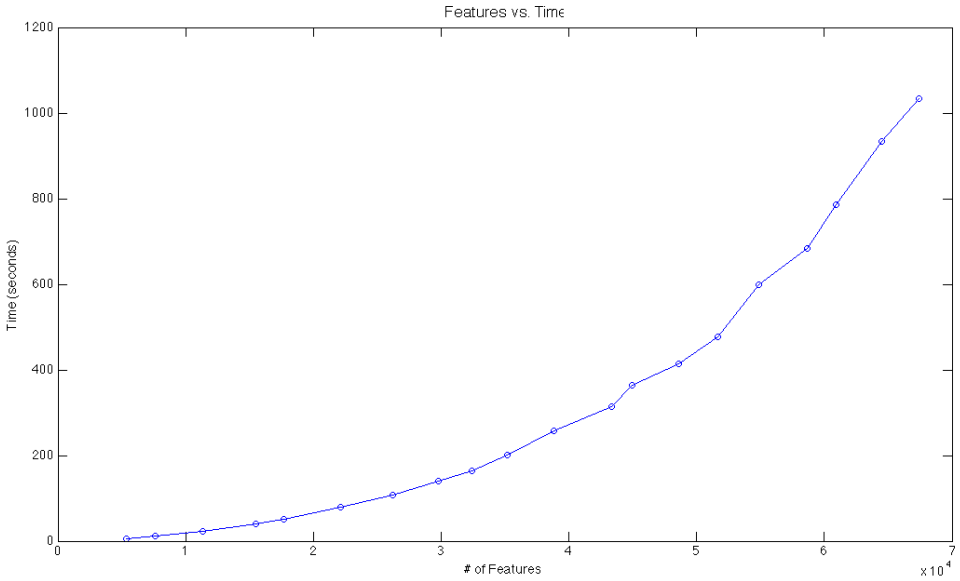


Figure 22 Classification Computation Time vs. # of Reviews of Venetian

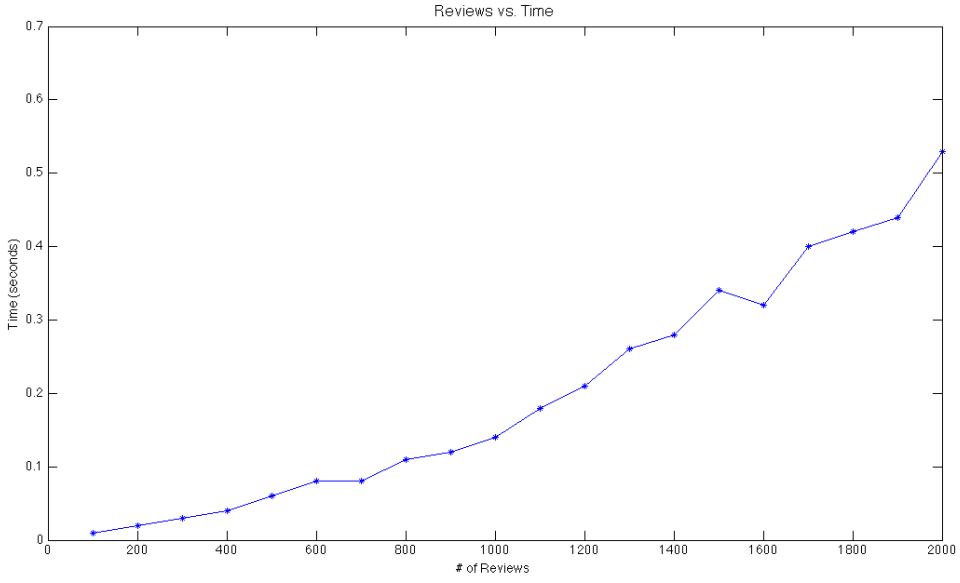


Figure 23 Classification Computation Time vs. # of Features of Venetian

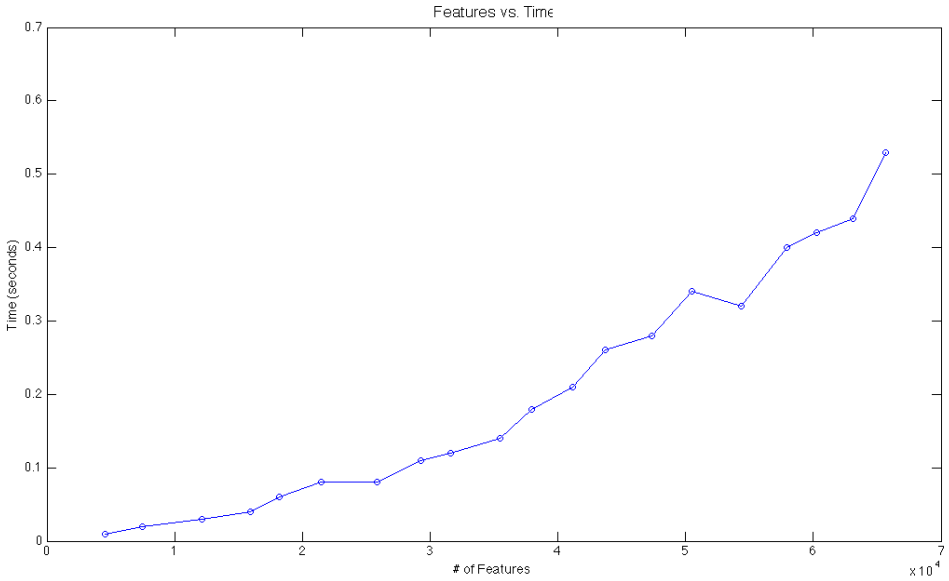


Table 6 2-star Americas Best Value Inn Descriptive Statistics

	All	Business	Couple	Family	Friend
Recommendation (actual)	59%	59%	74%	45%	53%
# of reviews	170	17	46	31	36

Table 7 3-star Bally Descriptive Statistics

	All	Business	Couple	Family	Friend
Recommendation (actual)	85%	79%	86%	81%	86%
# of reviews	2004	226	831	350	390

Table 8 4-star Treasure Island Descriptive Statistics

	All	Business	Couple	Family	Friend
Recommendation (actual)	85%	83%	86%	82%	85%
# of reviews	2858	251	1332	519	569

Table 9 5-star Venetian Descriptive Statistics

	All	Business	Couple	Family	Friend
Recommendation (actual)	92%	92%	92%	87%	92%
# of reviews	2025	188	1109	296	342

Table 10 Prediction Results of Bally (Accuracy)

Number of Features	All	Business	Couple	Family	Friend
10	85.97%	82.47%	87.61%	83.71%	88.69%
25	87.12%	86.73%	88.69%	85.71%	89.46%
50	88.72%	88.94%	90.25%	85.71%	90.49%
75	88.97%	87.17%	90.49%	87.14%	90.23%
100	89.57%	88.05%	90.37%	87.71%	91.00%
150	89.72%	88.05%	91.10%	87.71%	90.75%
200	90.16%	88.05%	90.73%	87.71%	90.23%
250	90.02%	87.17%	90.85%	87.14%	88.95%
300	90.41%	88.05%	90.97%	86.57%	88.69%
350	90.61%	88.05%	90.73%	87.14%	88.43%
400	90.61%	88.05%	90.85%	86.29%	87.40%
1000	90.56%	87.17%	89.77%	84.57%	85.86%
5000	90.26%	82.74%	86.16%	82.57%	86.12%
Chance criteria	74.5%	66.82%	75.92%	69.22%	75.92%

Table 11 Prediction Results of Bally (F-measure)

Number of Features	All	Business	Couple	Family	Friend
10	0.807	0.797	0.837	0.789	0.858
25	0.838	0.842	0.862	0.827	0.873
50	0.868	0.874	0.889	0.827	0.887
75	0.873	0.849	0.894	0.847	0.879
100	0.881	0.861	0.892	0.857	0.891
150	0.884	0.861	0.9	0.857	0.887
200	0.889	0.861	0.896	0.859	0.879
250	0.889	0.849	0.896	0.852	0.857
300	0.893	0.861	0.897	0.844	0.853
350	0.895	0.861	0.893	0.849	0.848
400	0.895	0.863	0.893	0.839	0.828
1000	0.895	0.853	0.872	0.823	0.793
5000	0.893	0.786	0.799	0.781	0.799

Table 12 Prediction Results of Bally (ROC)

Number of Features	All	Business	Couple	Family	Friend
10	0.545	0.632	0.57	0.58	0.615
25	0.602	0.688	0.627	0.639	0.65
50	0.667	0.74	0.698	0.639	0.679
75	0.681	0.698	0.714	0.671	0.655
100	0.7	0.719	0.709	0.692	0.682
150	0.71	0.719	0.724	0.692	0.673
200	0.718	0.719	0.715	0.698	0.655
250	0.722	0.698	0.709	0.688	0.609
300	0.729	0.719	0.709	0.673	0.6
350	0.728	0.719	0.701	0.677	0.591
400	0.728	0.726	0.694	0.665	0.555
1000	0.731	0.713	0.637	0.649	0.5
5000	0.733	0.609	0.504	0.573	0.509

Table 13 Prediction Results of Treasure Island (Accuracy)

Number of Features	All	Business	Couple	Family	Friend
10	86.07%	85.66%	86.38%	85.66%	87.35%
25	86.88%	89.64%	87.44%	86.82%	88.75%
50	87.86%	90.84%	88.20%	87.60%	90.16%
75	88.03%	90.44%	88.88%	86.24%	89.81%
100	88.52%	90.44%	90.09%	86.63%	90.16%
150	89.54%	90.04%	90.32%	87.21%	91.21%
200	89.96%	90.84%	91.45%	87.21%	91.21%
250	90.31%	90.44%	91.75%	87.21%	91.39%
300	90.27%	89.24%	91.60%	87.40%	91.56%
350	90.41%	89.24%	91.30%	87.40%	91.56%
400	90.69%	88.45%	91.38%	87.21%	91.56%
1000	90.87%	88.05%	89.79%	87.21%	91.04%
5000	89.89%	84.86%	86.54%	85.85%	88.75%
Chance criteria	74.5%	71.78%	75.92%	70.48%	74.52%

Table 14 Prediction Results of Treasure Island (F-measure)

Number of Features	All	Business	Couple	Family	Friend
10	0.813	0.826	0.805	0.823	0.843
25	0.837	0.883	0.839	0.844	0.868
50	0.856	0.897	0.856	0.86	0.885
75	0.861	0.894	0.863	0.842	0.883
100	0.867	0.894	0.885	0.844	0.886
150	0.884	0.888	0.891	0.851	0.902
200	0.888	0.896	0.903	0.852	0.903
250	0.892	0.892	0.908	0.851	0.905
300	0.892	0.878	0.905	0.854	0.907
350	0.895	0.878	0.9	0.855	0.908
400	0.898	0.866	0.9	0.852	0.908
1000	0.898	0.861	0.873	0.856	0.9
5000	0.889	0.813	0.806	0.835	0.87

Table 15 Prediction Results of Treasure Island (ROC)

Number of Features	All	Business	Couple	Family	Friend
10	0.547	0.618	0.51	0.623	0.606
25	0.596	0.725	0.574	0.663	0.662
50	0.637	0.751	0.613	0.701	0.69
75	0.653	0.749	0.622	0.668	0.698
100	0.664	0.749	0.679	0.666	0.695
150	0.708	0.737	0.701	0.678	0.74
200	0.717	0.742	0.724	0.682	0.75
250	0.721	0.739	0.74	0.678	0.705
300	0.722	0.714	0.727	0.683	0.756
350	0.736	0.714	0.709	0.687	0.761
400	0.739	0.69	0.708	0.682	0.761
1000	0.732	0.679	0.638	0.694	0.739
5000	0.722	0.595	0.511	0.653	0.672

Table 16 Prediction Results of Venetian (Accuracy)

Number of Features	All	Business	Couple	Family	Friend
10	91.60%	93.09%	92.70%	87.50%	91.98%
25	92.30%	93.09%	93.24%	89.53%	92.69%
50	93.19%	94.15%	94.23%	92.91%	94.15%
75	93.33%	93.09%	94.05%	92.23%	93.86%
100	93.18%	92.55%	93.42%	91.55%	93.57%
150	93.53%	92.55%	94.59%	90.54%	92.98%
200	94.07%	92.55%	94.77%	88.85%	93.27%
250	94.33%	92.55%	94.86%	88.85%	92.69%
300	94.37%	92.02%	94.59%	88.85%	92.40%
350	94.37%	92.02%	94.23%	88.18%	92.40%
400	94.42%	92.02%	94.41%	87.50%	92.11%
1000	92.94%	92.02%	92.88%	87.16%	91.81%
5000	91.70%	92.02%	92.36%	87.16%	91.81%
Chance criteria	85.28%	85.28%	85.28%	77.38%	85.28%

Table 17 Prediction Results of Venetian (F-measure)

Number of Features	All	Business	Couple	Family	Friend
10	0.876	0.906	0.895	0.84	0.911
25	0.895	0.917	0.91	0.869	0.906
50	0.916	0.925	0.931	0.917	0.928
75	0.918	0.906	0.931	0.907	0.924
100	0.917	0.894	0.921	0.897	0.919
150	0.925	0.894	0.936	0.88	0.908
200	0.931	0.894	0.939	0.849	0.911
250	0.935	0.894	0.939	0.849	0.899
300	0.956	0.882	0.936	0.849	0.893
350	0.936	0.882	0.931	0.855	0.893
400	0.936	0.882	0.932	0.82	0.886
1000	0.911	0.882	0.901	0.812	0.879
5000	0.878	0.882	0.887	0.818	0.879

Table 18 Prediction Results of Venetian (ROC)

Number of Features	All	Business	Couple	Family	Friend
10	0.5	0.567	0.524	0.558	0.604
25	0.552	0.628	0.575	0.615	0.586
50	0.626	0.633	0.656	0.724	0.659
75	0.632	0.567	0.676	0.697	0.641
100	0.634	0.533	0.63	0.671	0.623
150	0.671	0.533	0.679	0.632	0.588
200	0.69	0.533	0.697	0.566	0.589
250	0.705	0.533	0.692	0.566	0.684
300	0.715	0.5	0.679	0.566	0.554
350	0.71	0.5	0.661	0.539	0.536
400	0.71	0.5	0.657	0.513	0.536
1000	0.606	0.5	0.541	0.5	0.518
5000	0.506	0.5	0.5	0.511	0.5

Table 19 Comparisons of 2-star vs. 5-star (importance)

TF-IDF	Common features	Unique features
2-star	location, room, friend, restaurant, service, pool, comfort	basic room conditions (bed, towel, clean, door, shower), value, price, cheap
5-star	room, service, location, friend, restaurant (higher), comfort (higher), help (higher), pool(higher)	additional service (shop, getaway, show, dining, casino), luxury

Table 20 Comparisons of 2-star vs. 5-star (discriminating)

Chi-square value	Common features	Unique features
2-star	location (higher), rude(higher), room, manager(higher), furniture(higher), carpet(higher), internet(higher), sheet(higher)	door, value, toilet, carpet, mold, staff-very--helpful
5-star	room (higher), service (higher), rude, carpet, manager, furniture internet, location, sheet	Getaway, stain

Table 21 Comparison across Segments of Bally (importance)

TF-IDF	Common features	Unique features
Business	location, room, service, value, coffee, staff, clean, comfort(higher), friend, pool, restaurant, casino, help, buffet	Conference, internet
Personal	location, room, service, value, coffee, staff, clean, friend (higher), comfort , restaurant, help (higher), casino(higher), pool(higher)	additional service (parking, show, shop, getaway, view)

Table 22 Comparison across Segments of Bally (discriminating)

Chi-square value	Common features	Unique features
Business	rude, staff	bathroom, dirty, tv, wall paper, worn, noise, towel, price, value, wall, mold
Personal	rude(higher), staff(higher)	room, location, manager, unfriendly, furniture, clean, floor, year-wedding-anniversary, couch, casino, microwave, getaway

Table 23 Comparison across Segments of Treasure Island (importance)

TF-IDF	Common features	Unique features
Business	location, room, service, value, coffee ,staff, clean, comfort, friend , pool, restaurant, casino, help	conference, internet, center, fitness
Personal	location, room, service, value, coffee, staff, clean, friend (higher), comfort , restaurant (higher), help (higher), casino(higher), pool(higher), buffet(higher)	additional service (show, getaway, view, siren, shop, drink, birthday, anniversary, weekend, gamble)

Table 24 Comparison across Segments of Treasure Island (discriminating)

Chi-square value	Common features	Unique features
Business	Staff-friendly, comfort, rude	noise, friend, service, price, poor customer-service, spa, fitness-center
Personal	Staff- very- unfriendly (higher), comfort(higher), rude(higher)	occasion, manager, staff, fountain, staff- not- helpful, pillow, service, lobby, buffet, getaway

Table 25 Comparison across Segments of Venetian (importance)

TF-IDF	Common features	Unique features
Business	room, service, location, restaurant, friend, comfort, staff, security, help, shop, casino, pool, view, show	conference, internet
Personal	rooms, service, location, restaurant, friend(higher), comfort(higher), staff, security, help(higher), shop(higher), casino(higher), pool(higher), view(higher), show(higher),	additional service (getaway, weekend, gamble, anniversary, romantic, birthday, atmosphere, lobby, breakfast), luxury

Table 26 Comparison across Segments of Venetian (discriminating)

Chi-square value	Common features	Unique features
Business	room	door, hallway, closet, restaurant, newspaper, service, rate
Personal	room	service, rude, carpet, getaway, stains

Table 27 Computing Times of Indexing and Classification for Bally

# of reviews	# of features	Indexing time (seconds)	Classification time (seconds)
100	3663	4.135	0.01
200	8219	11.485	0.03
300	10324	17.028	0.04
400	12541	25.463	0.05
500	16917	43.278	0.07
600	19871	59.991	0.08
700	22698	75.989	0.08
800	25203	97.003	0.13
900	28474	122.071	0.15
1000	32097	153.097	0.19
1100	34080	180.335	0.23
1200	37624	227.832	0.24
1300	39888	258.586	0.3
1400	42719	301.9	0.31
1500	46267	365.205	0.33
1600	48571	379.345	0.45
1700	51638	465.237	0.49
1800	54068	539.049	0.56
1900	56468	629.5	0.62
2000	61568	739.902	0.71

Table 28 Computing Times of Indexing and Classification for Treasure Island

# of reviews	# of features	Indexing time (seconds)	Classification time (seconds)
100	4563	4.693	0.01
200	7504	10.253	0.02
300	12160	22.993	0.04
400	15920	38.172	0.04
500	18207	49.84	0.05
600	21485	67.808	0.08
700	25862	94.314	0.1
800	29284	136.592	0.16
900	31618	164.375	0.13
1000	35468	209.664	0.13
1100	37937	241.075	0.22
1200	41139	298.131	0.25
1300	43737	350.787	0.29
1400	47338	415.293	0.31
1500	50489	462.153	0.35
1600	54378	510.092	0.39
1700	57945	679.583	0.47
1800	60242	787.943	0.44
1900	63090	900.307	0.58
2000	65692	1028.843	0.63

Table 29 Computing Times of Indexing and Classification for Venetian

# of reviews	# of features	Indexing time (seconds)	Classification time (seconds)
100	5341	6.188	0.01
200	7654	10.953	0.02
300	11353	21.819	0.03
400	15497	40.317	0.04
500	17711	51.797	0.06
600	22107	78.462	0.08
700	26224	106.884	0.08
800	29815	139.602	0.11
900	32399	165.214	0.12
1000	35200	200.307	0.14
1100	38838	257.283	0.18
1200	43356	313.543	0.21
1300	44919	364.888	0.26
1400	48594	414.205	0.28
1500	51675	477.706	0.34
1600	54859	599.866	0.32
1700	58649	684.569	0.4
1800	60918	785.483	0.42
1900	64467	934.292	0.44
2000	67429	1033.645	0.53

Table 30 Indexing Computation Times Comparison between Serial Algorithm (Single Core) and Parallel Algorithm (Eight Cores)

	Serial Algorithm Single Core (Seconds)	Parallel Algorithm Eight Core (Seconds)
America Inn	3.107	0.600
Bally	411.486	70.908
Treasure Island	880.649	162.636
Venetian	428.526	89.218

Chapter/Essay 3

How to Use Multimedia Search Trends to Predict Auto Sales

3.1. Introduction

Marketers are always interested in predicting market sales so that they can arrange firm activities accordingly. With this knowledge of predicted sales, marketing managers can make strategic decisions concerning various marketing activities such as whether to increase or decrease production levels, whether to change size of the sales force and whether to initiate a price change to react to sales change trends and so on.

Traditional business or economic forecasting has relied on statistics collected by various agencies including government (like US Census Bureau) and business firms (like Automotive News Data center). However, there are always certain periods of delay of associated with such published data, which limited use of the forecasting especially for time-sensitive issues. These reports are usually available about half way through the next month and revised even several months later. There has also been very rich research in the marketing area on predicting market sales using marketing variables. One of the major areas is the use of purchase intention data to predict market sales (Morwitz, Steckel & Gupta, 2007; Kumar, Nagpal & Venkatesan, 2002; Amstrong, Morwitz & Kumar, 2000). Again there are issues regarding this predicting method. Traditionally intention data was collected through surveys and survey methods often suffer from bias (Podsakoff et al., 2003). In addition, surveying itself is often associated with limited number of observations obtained at a relatively high cost. Also in the past several decades, much of

the forecasting research has focused on complex mathematical models.

It would help marketers if they could find an easy and time efficient way to predict market sales. With the widespread adoption of search engines, people perform all kinds of search over the search engines such as Google all the time. Using the United States as an example, we can note that Americans performed 14.3 billion Internet searches in March 2009, which is an annualized rate of over 170 billion searches per year. Worldwide searches grew by 41% between 2008 and 2009. The large scale of online search activities helps us to gain an understanding of consumers' potential purchase intentions more accurately and timely. For example, Google trends provide daily, weekly and monthly search volume reports on various industries. With the help of related information technologies, near-real-time collection of search data can be obtained and at nearly zero cost. Each time a consumer performs a search of a product via the Internet, that individual's potential intention to purchase in the near future is revealed. In turn, these intentions can be used to predict future market sales like the old literature but on a larger scale, with higher accuracy (Choi and Varian, 2009; Shimshoni, Efron & Yossi, 2009; Wu & Brynjolfsson, 2014).

In this paper we attempt to predict automobile sales with Internet query data. Since text, image and video are the three major ways of information search, we included all these three types of search data as our predictors. We use the automobile market for our empirical analysis. We started making automobile market sales predictions from January of 2008 to January of 2013 and showed a very high accuracy. The rest of the paper is organized as follows: first we reviewed the related literature on sales prediction and using

search trends to predict sales. In section 3.3 we developed our methodology and applied it with empirical evaluations. We then reported the results and discussions in section 3.4. Finally we conclude the article with potential managerial applications and future research directions.

3.2. Literature Review

In the past decades, lots of social science work focused on applying complicated models to predict economic and social trends such as market sales. Nowadays, with the wide availability of the Internet and advanced information technology like online search and multimedia usage, the game has changed remarkably. Recently, a published book by James Surowiecki "The Wisdom of Crowds" pointed out the idea that a crowd of ordinary people can lead to the right decisions. A large scale of non-experts together, is even smarter than the experts. With the development of Internet and Information technology, lots of people are performing online search all the time and this real time data can be collected almost instantaneously and used to predict macro trends. Kuruzovich et al. (2008) showed that online behaviors could be used to reveal consumers' intentions and make predictions of purchases. Ginsberg et al. (2009) further showed how the use of Google trends and data from Centers for Disease Control (CDC) together find 45 specific search terms related to flu outbreaks and monitored influenza rates 1-2 weeks ahead of CDC reports.

Our work follows a similar stream in utilizing free and publicly available data from the Internet particularly from Google, to predict market sales. Choi and Varian (2009)

forecasted auto trends in the US using search queries from Google. Later Scott and Varian (2013) predicted some current economic trends like current gun sales using Bayesian variable selection techniques with search frequencies. Du and Kamakura (2012) demonstrated how to use a structural dynamic factor-analytic model to do quantitative trend spotting applied on Google trends data of automobiles. Wu and Brynjolfsson (2014) predicted future economic trends such as future sales in the housing market using trends data and government reports. However they have all looked at search queries of text information.

To the best of our knowledge, we are the first to combine the search of multimedia including web (mainly text), image and video (from YouTube) all together to predict future economic trends such as automobile sales. As we pointed out in the previous paragraphs, more search volume means more interested consumers are searching. So more search volume activity should lead to a higher level of sales. Text information can provide the richest information regarding every aspect of the product and is the most popular type of search. Image search can mainly provide the appearance of the searched product while a YouTube search of video can give a more in-depth experience of the car by looking at moving pictures with sound. Obviously the amount of search on image and video will be much less compared to the web search of text. We standardized the search volume index between 0-100 for comparisons (see data part for details). Later we normalized the search volume data of the three types to get the standardized estimation.

Text information over the web contains all the detailed information regarding the product from various websites like the official website and various dealer websites. This is the

most effective search method since people can get all different types of information with just one search. Image search can only list out the image of the product and lots of time the search is done for fun to look at but maybe less not for serious buying purposes. The impact of image on sales should be quite small compared to the other two types of search. Information search of video requires more effort than searching text or image like additional waiting time to load the video. People must have serious interest in the product to initiate a search. And most of the videos are commercial videos which are very persuasive to help lead to a sale or at least build a good brand impression. Based on the above, we come up with our Hypotheses regarding the relationship between the three types of search and market sales.

H1: There is a positive relationship between search volume of text and market sales

H2: There is a positive relationship between search volume of image and market sales

H3: There is a positive relationship between search volume of video and market sales

H4: The relationships between search and sales are moderated by origin of the car and type of the car.

3.3. Data and Model

3.3.1. Data

3.3.1.1. Google Trends Data

AC Nielsen NetRatings consistently placed Google as the top search engine across the world. Google processed more than 66.7% of all the online search queries in the world in December 2012 (comScore 2012). Especially in America, the search queries submitted in Google can reflect a large portion of Americans' intentions and interests. That is why we choose to use Google Trends to obtain the search data. We collected the search volume index related to the automobile category from the Google trends. The search volume index of each key search term (for example "Honda") is a relative share instead of the absolute number of queries submitted in Google trends by people. The search volume index for each query is defined as following: the total query search volume in a given geographic region at a given point of time divided by the total number of queries searched in that region at a given point of time. The query index is usually calculated at a weekly or monthly level depends on how popular the search is. So the search trends data is always between 0 and 100.

Google trends has also categorized the submitted queries into several predefined categories such as auto and vehicle (used in our study), home and appliance, computer and electronics and so on. When you perform the query search, Google trends also allows you to choose geographic regions such as United States, Canada, worldwide or other countries. You can also choose the period of time the query is submitted and the type of

the query submitted, whether it is text search on the web, image search or video search by YouTube. In today's multimedia era, people typically perform online search from text, image and/or video aspects and that's why we choose theses three as the predictors. We have extracted the time range of (2008/01-2013/01) and this consists of 61 data periods.

Figure 24, Figure 25 and Figure 26 show examples of the search query index of "Honda" by text from the web, image and video of YouTube in United States from 2008/01 to 2013/01 under auto and vehicles category.

----------Insert Figure 24, Figure 25 and Figure 26 here----------

3.3.1.2. Automobile Sales Data

We collected automobile sales data from Auto news data center for 36 make sold in the U.S. from 2008/01 to 2013/01.

Figure 27 shows an example of auto sales for Honda in the U.S. from 2008/01 to 2013/01.

----------Insert Figure 27 here----------

3.3.2. Model

Forecasting is widely used in marketing and economics. Traditionally, researchers assume the time series data to follow either a stationary stochastic process such as reflected in autoregressive (AR) model and autoregressive moving average process (ARMA) model or non-stationary process such as Fuller tests. In our study, we have a panel of auto sales data across different makes of automobiles and over a period of time.

Since we have reasonable sized T (time period) and not very large N (number of observation), the data falls into the TSCS data category (Beck & Katz, 1995). TSCS data has become popular more recently especially in the study of politic science; studies have considered for example, situation where the number of democratic countries is limited to 15 observed over a long period of time (Stimson, 1985; Alvarez et al., 1991; Maoz & Russett, 1993). This is very similar to our data with limited 36 makes and 61 periods. Terefore we applied the estimation methods for TSCS data.

The Generic model we consider has the form:

$$y_{it} = \sum_{k=1}^{K} x_{itk}\beta_k + u_{it} \quad i = 1, \ldots N; \quad t = 1, \ldots T_i$$

with the specification of u_{it} dependent on the particular model. The total number of observations $M = \sum_{i=1}^{N} T_i$. The $M \times M$ covariance matrix of u_{it} is denoted by **V**. Let **X** and **y** be the independent and dependent variables arranged by cross section and by time within each cross section. Let $\mathbf{X}_s$ be the X matrix without the intercept.

For TSCS data estimation, there are several common estimation methods including random effects, fixed effects and parks methods. In our case, we have a small number of cross sectional data over a long period of 61 months for all units. Parks method best fits our need since it deals with error complications by specifying respectively the error structure for heteroskedasticity, contemporaneous and serial correlations.

Parks Method (Autoregressive Model)

Parks (1967) considered the first-order autoregressive model in which the random errors u_{it} , $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, have the structure

$$\begin{aligned} E(u_{it}^2) &= \sigma_{ii} && \text{(heteroscedasticity)} \\ E(u_{it}\,u_{jt}) &= \sigma_{ij} && \text{(contemporaneously correlated)} \\ u_{it} &= \rho_i u_{i,t-1} + \epsilon_{it} && \text{(autoregression)} \end{aligned}$$

The model assumed is first-order autoregressive with contemporaneous correlation between cross sections. β is then estimated by generalized least squares.

3.3.3. Model Evaluations

The conventional R-squared measure is inappropriate for all models that the TSCS procedure estimates since a number outside the 0-to-1 range may be produced. Hence, a generalization of the R-squared measure is reported by Buse 1973 and this adjust R-square is used in our study for model evaluation.

3.4. Empirical Evaluations

In our study, we focused on using three major types of search including text, image and video over the biggest search engine – Google to predict the market sales of automobiles. Since we have multiple make of automobiles over the same period of time, we applied TSCS estimation methods.

The general model to predict market sales is like following:

$Sales_{it} = Intercept + TextSearch_{it} + ImageSearch_{it} + VideoSearch_{it} + u_{it}$

We applied the Parks estimation method to predict the sales of automobiles. Table 31, Table 32, Table 33 and Table 34 show the estimation results.

We can see generally, the model fits very well with R square falls in the range of 0.12 – 0.42. Now let's looks at the detailed estimation results of coefficients.

First in the Table 31 and Table 32, estimation of the overall data including all makes and segmented data of makes from US, Europe and Asia are listed. Overall, it fits our hypotheses. The search of video over YouTube has the biggest explanatory power of market sales with coefficient 0.13 over text (0.05) and Image (0.03) for the whole dataset. The origin of the make moderated the effect. Video search has the biggest impact on explaining car sales for car makes from the U.S. with a coefficient of 0.11. This is consistent with the major characteristics associated with the marketing of the US cars: they highly rely on video commercials for their advertising and best represented by videos. Search of text over the web has the biggest impact on prediction of car sales for car makes from Europe with a coefficient of 0.16. This is consistent with the major characteristics of the European cars: high quality with superior benefits, which can be best explained with text. People who search for European cars would love to get more detailed information of the car from the text information. For cars originating from Asia, all three types of search have similar impact on predictions of sales with coefficient 0.03, 0.07 and 0.05. There is no particular preference among the three search types.

----------Insert Table 31 and Table 32 here----------

Next we compare luxury car and non-premium cars.

Next we compare luxury cars and normal cars. For luxury cars (including makes like Acura, Audi, BMW…) consumers tend to use image and video to explore the information regarding the car and enjoy the hedonic feelings brought by the luxury cars; so search of image and video should best explain the sales of luxury cars with coefficients of 0.42 and 0.57. For non-premium cars (including makes like Honda, Toyota, Chevrolet…), people try to find specification details of the car from various channels to get a full understanding before they make purchase. Search of text over web (coefficient of 0.1) with search of image (coefficient of 0.04) and video (coefficient of 0.05) together can explain market sales since no particular type of search is preferred.

----------Insert Table 33 and Table 34 here----------

3.5. Conclusions and Future Research

With the development of technology, Internet search has become more and more popular. Consumers search over the Internet to get information with the purchase intention in heart before purchase. With the help of Google trends, we can literally get over billions of consumers' search data at near zero cost immediately. With the help of online search data, we can predict market sales more timely and accurately.

This is an exploratory research. We successfully use three types of search data: text, image and video to predict market sales of automobiles. This can benefit the auto industries significantly. Buying an automobile is one of the major expenditures for most

people. With accurate and timely predictions of sales, automobile companies can make better make managerial decisions. They can prepare well for future production scheduling, marketing, sales, and inventory planning activities. Depending on the origin of the car, marketers can focus on the most appropriate way to present information to potential consumers. For example, for US car makes, the most important way to communicate with potential consumers will be through videos and they should monitor the trends of videos timely to discover new trends. For Europe originated cars, website construction with ample text information is necessary and they should pay special attention to the search trends on text. For Asian car makes, no particular emphasis is there so they should pay attention to all of the three equally. We also found that for luxury cars, videos and images are definitely the best way to present the features of the cars. In the future, we can also consider the hedonic and utilitarian aspects of the various cars to see if the impact of different search types on explaining sales is different.

This method of prediction can be used not only for automobile industries but also in various other industries to predict future sales using Google trends data. Instead of paying a premium for industry reports or waiting for delayed government reports, Google trends can help us. Search data can also be combined with other types of data like text reviews on websites, for predicting not only sales but also word of mouth patterns.

However, this method also has certain limitations. For example, as the popularities of online search query become more and more, there might be manipulations of the search volume index due to usage of false or misleading search queries; such wrongly generated search volume index values could lead biased results affecting the market in many

adverse ways. Research on how to detect manipulated search queries may become a new research direction in the future.

3.6. References

Alvarez, R. M., Geoffrey G. & Lange, P. (1991). "Government partisanship, labor organization and macroeconomic performance". American Political Science Review, 85(2):539–556.

Armstrong, J. S., Morwitz, V. G. & Kumar, V. (2000). "Sales forecasts for existing consumer products and services: Do purchase intentions contribute to accuracy". International Journal of Forecasting, 16(3):383−397.

Beck, N & Katz, J. N. (1995). "What To Do (and Not To Do) with Time- Series Cross-Section Data". American Political Science Review, 89(3):634–647.

Buse, A. (1973). "Goodness of Fit in Generalized Least Squares Estimation". American Statistician, 27(3):106-108.

Choi, H. & Varian, H. R. (2009). "Predicting the present with Google trends". Economic Record, 88(1);2-9.

Du, R. & Kamakura W. A. (2012). "Quantitative Trendspotting". Journal of Marketing Research, 49(4):514-536.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2008). "Detecting influenza epidemics using search engine query data". Nature, 457:1012-1014.

Kumar, V., Nagpal, A. & Venkatesan, R. (2002). "Forecasting category sales and market share for wireless telephone subscribers: A combined approach". International Journal of Forecasting, 18(4):583−603.

Kuruzovich, J., Viswanathan, S., Agarwal, R., Gosain, S. & Weitzman, S. (2008). "Marketspace or marketplace? online information search and channel outcomes in auto retailing". Information Systems Research, 19(2):182-201.

Maoz, Z., & Russett, B. (1993). "Normative and structural causes of democratic peace, 1946-1986". American Political Science Review, 87(3):624-638.

Morwitz, V. G., Steckel, J. H. & Gupta, A. (2007). "When do purchase intentions predict sales". International Journal of Forecasting, 23(3):347–364.

Parks, R. W. (1967). "Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated". Journal of the American Statistical Association, 62(318):500-509.

Podsakoff , P., MacKenzie, S., Lee, J. & Podsakoff, N. (2003). "Common method biases

in behavioral research: A critical review of the literature and recommended remedies". Journal of Applied Psychology, 88(5):879–903.

Scott, S. L. & Varian, H. R. (2013). "Bayesian variable selection for nowcasting economic time series". National Bureau of Economic Research.

Shimshoni, Y., Efron, N. & Matias, Y. (2009). Google Israel Lab.

Stimson, J. A. (1985). "Regression in space and time: A statistical essay". American Journal of Political Science, 29(4):914-947.

Surowiecki, J. (2005). The Wisdom of Crowds. Random House LLC.

Wu, L. & Brynjolfsson, E. (2013). "The future of prediction: How Google searches foreshadow housing prices and sales". National Bureau of Economic Research.

Figure 24 Search Index of text for "Honda"

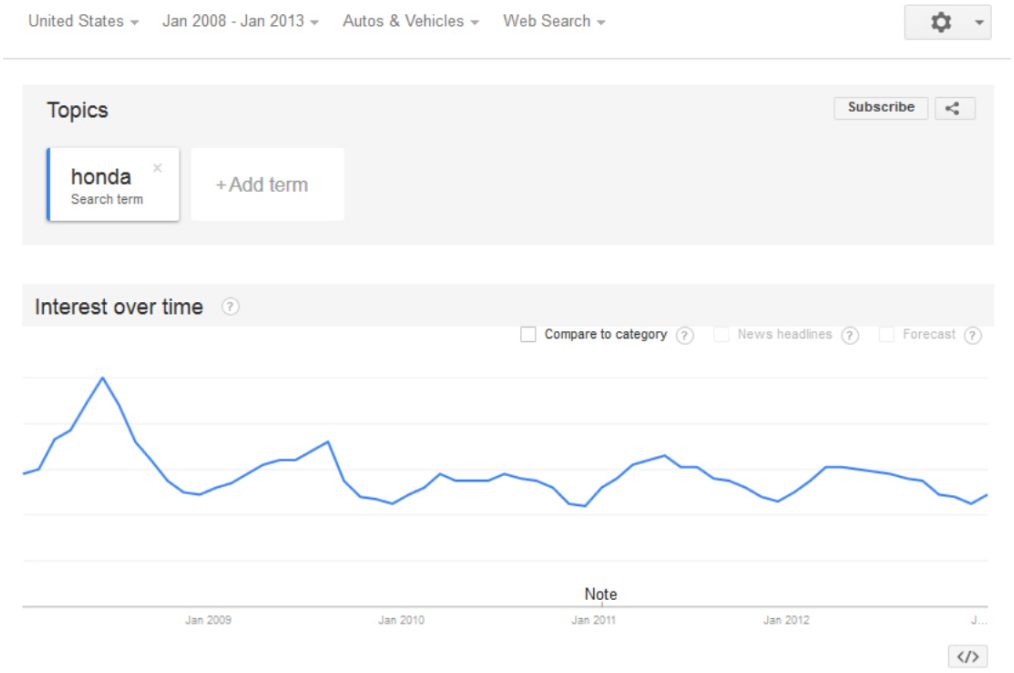


Figure 25 Search Index of image for "Honda"

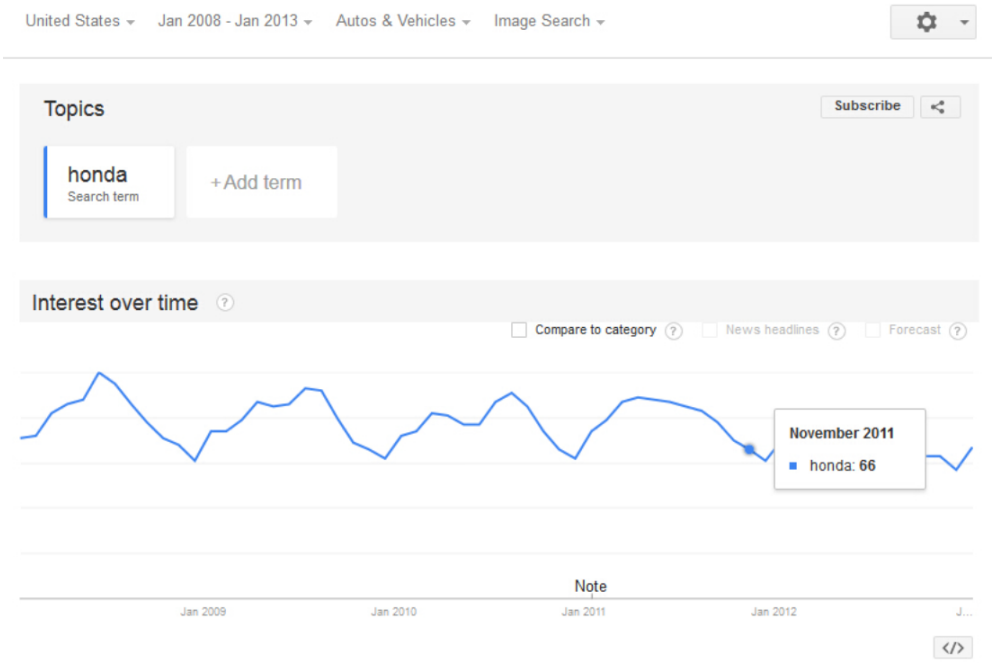


Figure 26 Search Index of video for "Honda"

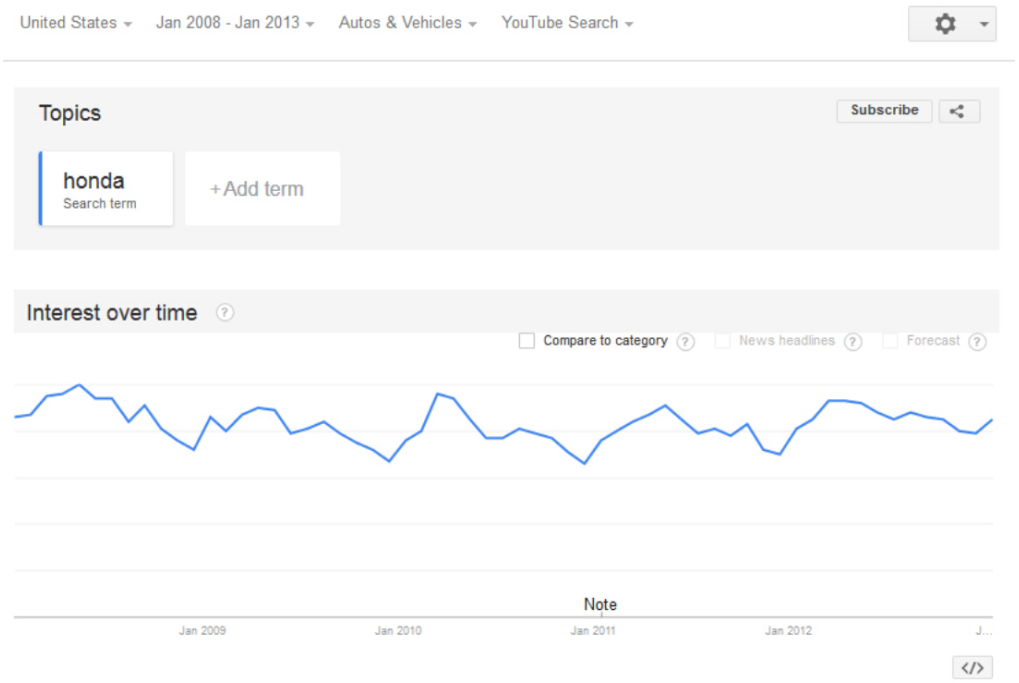


Figure 27 Auto sales for Honda in US

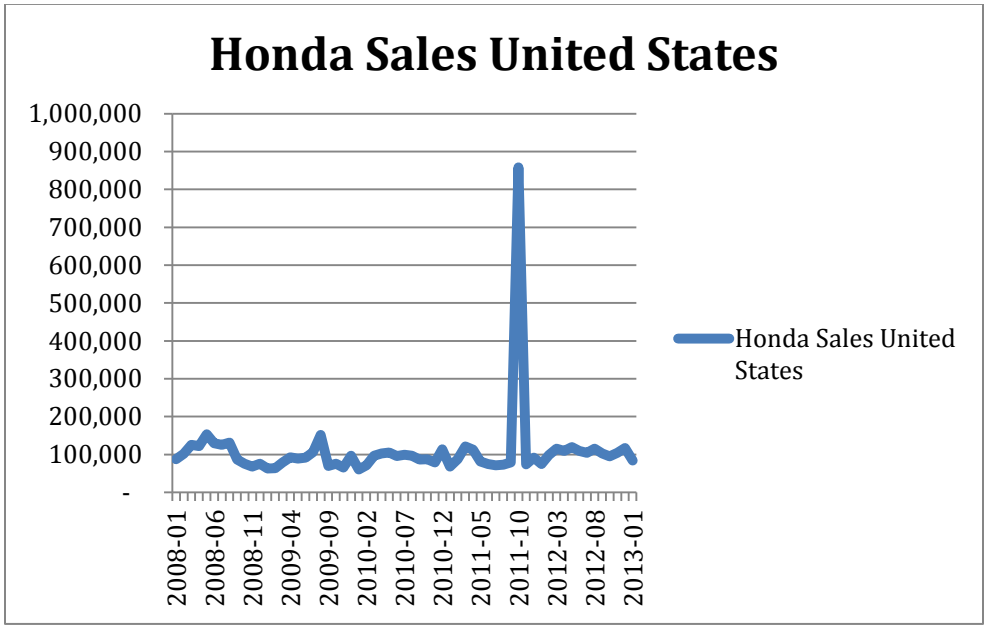


Table 31 Model Comparisons for Total Make and Different Origins

	Total Make	US	Europe	Asia
R-square	0.4211	0.1229	0.1358	0.2057

Table 32 Coefficient Estimations for Total Make and Different Origins

	Total Make	US	Europe	Asia
	-0.15 (<0.0001)*	-0.33 (<0.0001)*	-0.29 (<0.0001)*	-0.27 (<0.0001)*
Text	0.05 (<0.0001)*	-0.02 (0.1900)	0.16 (<0.0001)*	0.03 (0.0039)*
Image	0.03 (<0.0001)*	-0.007 (0.6486)	-0.04 (<0.0011)*	0.07 (<0.0001)*
Video	0.13 (<0.0001)*	0.11 (<0.0001)*	0.001 (0.8211)	0.05 (<0.0001)*

Table 33 Model Comparisons for Luxury VS. Non-premium Car

	Luxury	Non-premium
R-square	0.1852	0.3660

Table 34 Coefficient Estimates for Luxury VS. Non-premium Car

	Luxury	Non-premium
Intercept		
Text	-0.02 (0.5225)	0.04 (<0.0001)*
Image	0.42 (<0.0001)*	0.05 (<0.0001)*
Video	0.57 (<0.0001)*	0.10 (<0.0001)*

CURRICULUM VITAE

Name: Shaoqiong Zhao
Dissertation Title: Three Essays on Consumers' Activities in the Online Domain
Place of birth: Shanxi, China

Education

Ph.D., University of Wisconsin-Milwaukee, August 2014
Major: Marketing Minor: Supply Chain & Operations Management

M.S. University of Wisconsin-Milwaukee, May 2012
Major: Marketing

B.A. Peking University, July 2006
Major: Pharmacy

Honors and Awards

Marketing Research Society, University of Wisconsin-Milwaukee

AMA-Sheth Foundation Doctoral Consortium Fellowship 2012

Sheldon B. Lubar Doctoral Scholarship 2009/2010/2011

Chancellor's Fellowship Award, University of Wisconsin - Milwaukee 2008

Excellent Graduate, Peking University 2006