The Impact of Sentiment Analysis on Decision Outcomes- An Empirical Investigation

Parisa Lak

MBA, Sharif University of Technology, 2009

B.Eng., Electrical Engineering, 2004

A thesis presented to Ryerson University in partial fulfillment of the requirements for the Degree of Master of Management Science in Management of Technology and Innovation

Toronto, ON, Dec. 2013

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

The Impact of Sentiment Analysis on Decision Outcomes- An Empirical Investigation

Parisa Lak

MBA, Sharif University of Technology, 2009

B.Eng., Electrical Engineering, 2004

A thesis presented to Ryerson University in partial fulfillment of the requirements for the Degree of Master of Management Science in Management of Technology and Innovation

Toronto, ON, Dec. 2013

Abstract

A typical trade-off in decision-making is between the cost of acquiring information and the decline in decision quality caused by insufficient information. Consumers regularly face this trade-off in purchase decisions. Online product/service reviews serve as sources of product/service related information. Meanwhile, modern technology has led to an abundance of such content, which makes it prohibitively costly (if possible at all) to exhaust all available information.

The Impact of Sentiment Analysis on Decision Outcomes- An Empirical Investigation

Consumers need to decide what subset of available information to use. Star ratings are excellent cues for this decision as they provide a quick indication of the tone of a review. However there are cases where such ratings are not available or detailed enough. Sentiment analysis –text analytic techniques that automatically detect the polarity of text– can help in these situations with more refined analysis.

This study was performed in two interrelated phases. In the first phase the potential impact of Sentiment Scores (sentiment analysis outcomes) was investigated through a comparison between these scores with an already established numerical rating denoted as star ratings in three different domains. The results show that sentiment scores tend to fall into neutral areas and are not able to detect extremes that were reported to be more beneficial for information acquisition purposes. As a result, to use the current sentiment analysis results as a substitute for star ratings, a partial linear filter was applied to sentiment analysis results in a way to highlight the subtle differences away from the "neutral zone".

In the second phase, the impact of the extended version of sentiment scores on decision outcomes was examined through a controlled experiment. The examined decision was a purchase decision and the information provided was pages of reviews annotated with extended sentiment scores on each paragraph. Human subjects were used in the experiment and controlled data gathering sessions was designed. Results suggest that female consumers may use sentiment scores on review documents without other comparison aids to increase their confidence level in their purchase decisions.

Acknowledgement

There are a number of people without whom this thesis might not have been written, and to whom I am greatly indebted.

I would like to express my appreciation to my supervisor, Professor Ozgur Turetken, for his great help and support during my research. Undoubtedly, without his guidance and persistent help this thesis would not have been possible.

I must acknowledge as well the many friends, families, colleagues and students who participated in my inquiries and assisted, advised, and supported my research.

I would like to thank the thesis committee members, Professor Aziz Guergachi, Professor Farid Shirazi, and Professor Ali Miri for taking their time to read this dissertation and give me valuable feedback during my defense meeting.

This research was supported by an NSERC discovery grant.

Dedication

This thesis is dedicated to my loving parents, Dr. Parviz Lak and Ms. Jaleh Zahed, for their love and endless support. Also, I would like to dedicate my work to my brother and sister, Pedram and Parinaz, whose words of encouragement always ring in my ears.

Table of Contents

Author's Declaration iii

Abstract iv

Acknowledgement vi

Dedication vii

Table of Contents viii

List of Tables xi

List of Figures xii

List of Appendices xiii

1. Introduction 1

1.1. Problem specification and research questions 3

1.2. Structure of the thesis 5

2. Literature Review 6

2.1. Consumer purchase decisions 6

2.2. Sentiment Analysis Technology 9

2.3. Sentiment Analysis Applications 12

3. Phase one- Star Ratings versus Sentiment Analysis - A Comparison of Explicit and Implicit Measures of Opinions 18

3.1. Introduction **18**

3.2. Methodology **18**

3.2.1. Data selection process 19

3.2.2. Sentiment Analysis Tool 20

3.3. Analysis of data **24**

3.4. Discussion and Conclusions **27**

4. Phase 2 – User Studies **30**

4.1. Introduction **30**

4.2. Theoretical background **31**

4.2.1. Cost/ benefit framework 32

4.2.2. Information Foraging theory 33

4.2.3. Task Technology Fit 36

4.3. Hypothesis and research model development **40**

4.4. Research methodology **43**

4.4.1. Experiment design 46

4.4.2. Measurement Scales 48

4.4.3. Ethics approval 49

4.4.4. Pilot study 49

4.4.5. Data collection 50

4.5. Data analysis and results **51**

4.5.1 Coding of variables 52

4.5.2. Descriptive statistics and frequencies of variables 55

4.5.3. Hypotheses Testing 59

5. Discussion and Conclusions **71**

6. Limitations 74

7. Directions for future research 76

8. Appendices 77

9. Bibliography 87

List of Tables

Table 1- Sentiment Analysis Applications 12

Table 2 - Lexalytics and Lymbix comparison 23

Table 3- Cross Tabulations of Normalized Sentiment Analysis Scores (SA) & Star Ratings (SR) 24

Table 4 - SPSS bivariate correlation results 27

Table 5 - Description of Variables and Coding 53

Table 6 - Factor analysis on Confidence factors 54

Table 7 - Descriptive statistics for Confidence 54

Table 8 - Descriptive Statistics for all Variables 55

Table 9 - Frequencies of Variables 56

Table 10 - Relationship between “Time to make decision” and “Confidence” 58

Table 11 - Distribution of use of sources of information 59

Table 12 - Test for the independency of covariates and treatment effect 61

Table 13 - Independent sample t-test on control variables 63

Table 14 - ANCOVA test for H_2 64

Table 15- ANCOVA to test $H_{3\text{-}1}$ 66

Table 16- ANCOVA to test H_3 on Males only 67

Table 17- ANCOVA to test H_3 on Females only 67

Table 18- Descriptive Statistics for H_3 69

List of Figures

Figure 3 - Task Technology Fit 38

Figure 4 - Proposed Model 42

Figure 5 - Original vs. Extended sentiment scores 46

Figure 6 - Distribution of Reading speed (left) and Time to make decision Variable (right) 57

Figure 7 – “time to make decision” versus “confidence” 58

Figure 8 - Test for normality using Q-Q plot 60

Figure 9 - Test for homogeneity of regression slopes 62

List of Appendices

Appendix 1- Power test 77

Appendix 2- Chi-square test results 78

Appendix 3- Consent Form 81

Appendix 4 - Questionnaire 85

1. Introduction

Market efficiency depends on the availability of information such as product specification and pricing. It is evident that decision makers benefit from more information while making a sale or purchase decision [62]. However, gathering information comes with a cost, which is the time and effort (and sometimes financial resources) spent to gather, analyze, and comprehend that information. Individuals are normally aware of the trade-offs between the perceived costs and benefits of search [27] and sometimes they add this cost to the final value of the product or services that they are going to purchase or use, and make decisions correspondingly.

One of the most important sources of product and service information is Word Of Mouth information. With the growing success of web 2.0 technologies and the emergence of social media interactions, user generated contents have produced a new version of WOM information. This information is produced each second in the form of tweets, blog posts, news, reviews, comments, etc. Among all these forms of information, reviews play a notably significant role in consumers' purchase decision making. The benefit of online reviews over the traditional WOM information is the accessibility and the variety of the information that can be gathered in less than a second.

Millions of people, nowadays, express their opinions about restaurants, hotels, products and even their family physicians or university professors through online review websites such as Yelp[1], Tripadvisor[2], Amazon[3], RateMds[4] and Ratemyprofessor[5]. This user-generated content can

[1] http://www.yelp.ca/

[2] http://www.tripadvisor.com/

[3] http://www.amazon.ca/

[4] http://www.ratemds.com/

[5] http://blog.ratemyprofessors.com/

be used by individuals to make wiser decisions [39]. However, this voluminous amount of information is sometimes hard for users to digest, and hence it gives rise to the challenge of information overload. To mitigate this challenge, information summarization or categorization techniques have been proposed using intelligent systems. These studies are presented under the umbrella of "big data management". One big data management technique with the promise of categorizing documents by their underlying emotions is "Sentiment Analysis".

Sentiment Analysis refers to a data mining and text analytics technique that detects the polarity of documents. Sentiment analysis tools use different classification techniques to determine whether a piece of text is positive, negative or neutral. The result will be presented in either a binary classification or with a specific quantitative measure. Some sentiment analysis tools are also able to express topic specific polarities as well as a general polarity score.

Various applications have been developed based on this basic principle. The introduction of this new technology in early 2000s has raised significant attention to this new area of research. Researchers have focused mainly on the design of a new system or the improvement of a currently available one to increase accuracy. However, most studies fail to provide support for the claim that document polarity detection is a useful text summarization technique that yields effective and efficient completion of different decision-making tasks.

As mentioned earlier, and will be discussed in more detail later in this thesis, there are various applications of sentiment analysis technology. Nonetheless, there is minimal, if any, attention to testing whether sentiment analysis technology is useful for those applications. Hence, in this study, we decided not to aim for designing a new system or to make any improvements to a currently available one, but to evaluate the technology's usefulness and its success in assisting

individuals perform a specific task. For this purpose we will use a publically available state of the art sentiment analysis system.

1.1. Problem specification and research questions

In this study, we claim that sentiment scores can be used as a clue to summarize documents and respectively improve information acquisition process, when other numerical ratings are not available. We assume that if successful information acquisition occurs, decision outcomes will be improved. We test this claim with a controlled experiment using human subjects. The decision that will be investigated is a purchase decision and the document being semantically scored is pages of product reviews. The outcome will be tested in terms of effectiveness and efficiency of the decision.

Before addressing the question of usefulness (of sentiment scores) to support our claim, there is a need to "validate" sentiment scores extracted from our selected system. The test for validation is designed not only to validate the numerical scores but also to explore the potential impact of these scores on decision outcomes. The sentiment analysis system, selected for this study, was chosen from a list of off the shelf, state of the art sentiment analysis systems (it will be explained further in this thesis). It can be considered an adequate representative of sentiment analysis systems, which delivers specific polarity scores rather than a binary report.

Since the methodology and approach to address the validation and examination of the potential impact of sentiment scores is different from the evaluation of its usefulness, we split our work into two interrelated phases. The first phase will address the validity and potential impact investigation, the result of which will initiate unique ideas to support the usefulness problem. In the second phase, we will evaluate the "usefulness" problem through an experiment that is designed to evaluate the impact of sentiment analysis on decision outcomes.

Validity of sentiment analysis will be evaluated by comparing sentiment scores for specific comments to their respective star ratings, which are common clues used by individuals to filter what they read during information acquisition. If we are able to support that sentiment scores are similar to star ratings, then it can be claimed that these numerical scores can be used interchangeably. The research question that we intend to address in this part of our study is:

Q1- How comparable are sentiment scores for reviews/comments to their respective star ratings?

It is expected that star ratings and sentiment scores are correlated if sentiment scores are valid. This will be true only if comments are consistent with their respective star ratings. If this semantic proximity between star ratings and sentiment analysis results can be established, then sentiment analysis measures can be considered surrogates for star ratings when such ratings are not available such as in long reviews, blog posts or news websites. This will also lead to extend the findings of research on the usefulness of star ratings to the usefulness of sentiment scores. Consequently, the result from this phase of our study will support the claim that sentiment scores can be used to facilitate decision processes in the same manner as star ratings.

In the second phase of our study, we will investigate how sentiment scores will impact individuals' purchase decisions while using online reviews as the only source of product information. The research question that we are planning to address in this phase is:

Q2- How do sentiment scores impact decision outcomes?

This question will be addressed by an empirical investigation with a controlled experiment using human subjects. Decision outcomes will be evaluated both objectively and subjectively. The objective measure for the decision will be the time that individuals spend to search, find and analyze the information that is provided to them and make their decision while the subjective measure will be assessed by user evaluation of their level of confidence about their decision.

The result from this phase will provide support for the claim that sentiment scores can be used as a clue to summarize documents and acquire an acceptable amount of information, which yields to a more efficient and effective decisions. We then may claim that these scores are useful for decision-making purposes, while information is presented in a long unstructured text document.

1.2. Structure of the thesis

The rest of this thesis is organized as follows. In the next chapter, section 2, we will do an extensive literature review on purchase decisions and word of mouth as well as sentiment analysis technology and its applications. As stated, this study was performed in two interrelated phases. Section 3 will cover details of the study on phase one. Phase 2 will be outlined in the next section (section 4). Discussion and conclusions will be presented in section 5. In the following section limitations of our study will be outlined and it will be followed by future research directions in section 7.

2. Literature Review

2.1. Consumer purchase decisions

Opinions are central to almost all human activities and are key influencers of our behaviors [1], hence by taking a close look at the opinions indicated by individuals we can predict certain behavior related to those opinions. Also, "our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world" [1]. Hence others' opinion plays a significant role in our decision making process.

Even in our minor decision making, we would like to hear others opinion and act accordingly, and that is what marketing specialists have been using as the basis for word of mouth marketing. Economic and marketing studies have extensively shown that word-of-mouth (WOM) plays an important role in shaping consumer attitudes and behaviors [2]. More specifically, it was proven by previous studies that purchase decisions are increasingly influenced by supplemental product information provided by user and consumer feedback [93, 94, 95].

It has been thought traditionally that the main reason consumers search for new information is to reduce their uncertainty about their decision [5, 6]. They will search for information until they reduce their uncertainty to a tolerable level [7]. This information can be obtained from different sources. Reviews, whether from a professional or a regular user, can be considered as one the best sources of product or service information. Online reviews, being accessible regardless of time and distance, can be deliberated for even better source of information than traditional paper base (flyers) reviews. Thus, buyers may use review websites to reduce their uncertainty about their decision [26].

However, the rapid increase in the volume of Internet users and the growth of web 2.0 (interactive web) popularity among those users, gave rise to massive collections of user-generated

content [46]. Hence, information finding, gathering, comprehending and use from this source of information have become more challenging.

Individuals normally are aware of the trade-offs between the perceived costs and benefits of search for more information [27]. Thus, they (often implicitly) calculate the total cost of a product as both the product cost and the cost of search for more information regarding that product [28]. For a wide range of choices, consumers recognize that there are tradeoffs between effort and accuracy [29]. Hence, information summarization and categorization techniques seem to be a good fit to reduce the effort and increase the accuracy.

On the other hand, "Among the many and varied channels through which a person may receive information, it is hard to imagine any that carry the credibility and, thus, the importance of interpersonal communication, or word of mouth (WOM)" [30]. With the extensive use of interactive web and that massive amount of user- generated content, online review websites have become one of the most useful sources of "word of mouth" information. Kumar and Benbasat [31] indicate that the presence of customer reviews on a website has been shown to improve customer perception of the usefulness and social presence of the website.

Review websites, mostly, require their users to rate products or services out of the scale of 5, denoted as star rating. Some websites give their users the opportunity to indicate their opinion by writing comments along with these rankings. Online consumer reviews are not exceptions to the rules of economics of information [32] in that, it is important to discern which reviews are the most useful and actually able to reduce consumers' purchase uncertainty. According to [33], star ratings provide an excellent opportunity to measure the valence of comments without analyzing the comments themselves.

Consumers can use decision and comparison aids [34] and numerical content ratings (such as star ratings) [35] not only to conserve cognitive resources and reduce energy expenditure to acquire information, but also to ease or improve the purchase decision process [26]. The star rating has been shown to serve as a cue for the review content [35].

There have been numerous studies on consumer's perception of usefulness of positive and negative reviews. For instance, in [36], Pavlou and Dimoka found that the extreme ratings (either 5 star or 1 star) of eBay sellers were more influential and useful than moderate ratings. Likewise, Forman et al. [37] found that for books, moderate reviews (3 stars) were less helpful than extreme reviews. However, Crowley and Hoyer [38] found that two-sided arguments (moderate reviews with 3 stars) are more persuasive than one-sided positive arguments when the initial attitude of the consumer is neutral or negative, but not in other situations.

Nevertheless, the utility of star ratings can be limited in certain contexts. For example, there are occasional reviews that are pages long yet with only an overall star rating assigned to the whole review. In such a case, the decision facing the consumer is regarding which part of the overall review to read. This is particularly relevant when comparing complex products and services with many features where it would be useful to have numeric scores for each specific feature separately.

Meanwhile, many other useful sources of reviews such as blog posts and news websites do not contain any numerical information resembling star ratings. Therefore the question arises as to which blog post or news website one should read given the limited (time) resources and the lack of additional cues such as star ratings on the products/services that these sources are reporting on. This is a "big data" problem as it is caused by not only the volume, but also the variety of data. One text analytics technology with promise to address this problem is sentiment analysis.

Sentiment Analysis is a technique that is mainly referred to as data summarization and opinion mining technique. Sentiment analysis, opinion mining, subjectivity analysis or review mining [15] are terms that are used interchangeably to refer to the process for detecting the polarity of a document. In their early work, Hatzivassiloglou and McKeown [9] reported that it is possible to identify sentiment words and their respective polarity in sentences with a high accuracy of 82%. This accuracy is reported by comparing the result of an automated system versus the analysis done by human. Following this finding, various Sentiment Analysis Algorithms have been proposed and numerous studies have been testing the accuracy of those algorithms [11, 16, 17, 18, 19, 20, 21, 22, 23].

2.2. Sentiment Analysis Technology

The year 2001 can be considered as the starting point of the widespread research on sentiment analysis and opinion mining [15]. This incident was majorly caused by the rise of machine learning methods, natural language processing and information acquisition methodologies. Also, the availability of data sets due to the blossoming of social web (web 2.0) as well as the realization of the intellectual challenges that the area offers were some other reasons that made researchers to be interested to this field [15].

System designers face numerous challenges designing sentiment analysis tools. Opinion extractions and classification is the main objective of sentiment analysis systems and hence defining subjective versus objective opinion is the biggest challenge that system designers encounter. There are certain questions that should be addressed before moving to the opinion extraction phase.

One main question is whether the system should express a binary value for positivity and negativity or an exact degree of positivity and negativity should be extracted. When ranking rather

(exact degree of polarity) is desired, certain categorization techniques or a combination of some should be used to define the exact polarity degree. For instance Niu et al. [102] used a supervised learning method to perform classification at the sentence level. They used various categorization techniques and tested performance for different combinations of feature sets to determine the polarity of outcomes described in medical text. They were able to find that combining linguistic features and domain knowledge leads to the highest accuracy to indicate medical text polarity in four different categories: no comment, positive comment, negative comment and neutral comment.

Another question to be answered is what type of classification techniques should be used? Some use Lexicon based algorithms for opinion extraction and classification while others use Machine Learning algorithms. Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM) are some [49], but not all, of various types of machine learning techniques used for sentiment extraction and classification.

The next step is the opinion extraction itself and the detection of subjectivity and objectivity of the text in general. Hatzivassiloglou and Wiebe [103] in their early work have found evidence that adjective orientation in documents is highly correlated with the subjectivity of that document. In that, if a positive adjective is detected in the document there is a high chance for the document to be positive. Moreover, Turney [104] proposed that selected phrases including an adjective or an adverb are better indicators of document subjectivity. Pang et al. in [11] also were able to find that some nouns and verbs are also playing a significant role in subjectivity detection of the document.

One other debate between system designers is whether to use term presence or term frequency. Pang and Lee [15] indicate, "While a topic is more likely to be emphasized by frequent occurrences of certain keywords, overall sentiment may not usually be highlighted through

repeated use of the same terms". Furthermore, Yang et al. [105] looked at rare terms that do not appear in any dictionary, but are used in blogs and carry subjective meanings and hence correlates with the subjectivity of the document.

A recent hot topic in opinion mining and text analytics is domain dependency. It is stated that words may have completely opposite meanings or different sentiment strengths from one domain to the other. Owsleys et al. [25] specify that in order to achieve the best sentiment analysis results a domain- specific lexicon should be built. This lexicon should be related to both the entities and their sentiment expressions. To build domain-specific lexicons, researchers have proposed various techniques. One common approach is to start from a small initial sentiment lexicon and gradually expand it during the processing of reviews, while another common approach is bootstrapping. Bootstrapping was introduced by Riloff and Wiebe [10] and is the approach for subjectivity classification that learns patterns of subjectivity clues from un-annotated texts. These clues will then be used by a Naive Bayes classifier to produce input for the pattern learner.

Additionally, most sentence level and even document level classification methods are based on identification of opinion words or phrases. There are basically two types of approaches for opinion words identification: (1) corpus-based approach, and (2) dictionary-based approach [11,12,13,14]. Corpus-based approaches find co-occurrence patterns of words to determine the sentiments of words or phrases, while dictionary-based approaches use synonyms and antonyms indicated in lexical English data bases such as WordNet and Epinion to determine word sentiments based on a set of opinion words.

As mentioned above various, extraction and classification techniques can be used to identify the tone of a given piece of documents. In general, sentiment Analysis systems perceive whether a text is positive, negative or neutral. This analysis can be aggregated over large sets of

data and the resulting information can be helpful in different contexts, which is going to be discussed in the following section.

2.3. Sentiment Analysis Applications

As stated above, Sentiment Analysis systems use extraction and classification techniques in order to indicate the polarity of the document. Various applications have been developed based on this basic principle. Some of those applications from recent literature are listed in table 1.

Table 1- Sentiment Analysis Applications

Author	Year	Sentiment Analysis applications	Potential outcomes
Huang et al. [39]	2013	▪ Decision making	▪ Make wiser decisions ▪ Make decisions significantly faster
Cambria et al. [41]	2013	▪ Information extraction	▪ Distilling useful information from unstructured data
Rosas et al. [42]	2013	▪ Branding and product analysis ▪ Tracking sentiment timelines in on-line forums and news ▪ Analysis of political debates ▪ Question answering ▪ Conversation summarization	
Paltoglu et al. [43]	2012	▪ Making Predictions ▪ Review summarization	▪ Estimates the level of emotional intensity contained in text in order to make a prediction

			▪ Aiming to predict whether a reviewer recommends a product or not
Liebmann et al. [44]	2013	▪ Finance and E- commerce	▪ Resource allocation of E-commerce ▪ Financial prediction (difference between analyst and investors decisions)
Zhao et al. [45]	2012	▪ User behavior evaluations	▪ Understanding user behaviors

It is noteworthy to mention that almost all publications on Sentiment Analysis present at least one application for this technology. Yet, most do not provide an empirical support for their claim. For instance, Huang et al. [39] indicate that sentiment analysis results can be used to make wiser decisions and to make those decisions significantly faster. Their argument is based on the idea that the extensive amount of product or service information extracted from user generated contents (mainly retrieved from web 2.0) is not easy to digest by individuals who are seeking to make a purchase decision using those information. This paper relies on the results from a previous study by Yatani et al. [96] on feature-sentiments information and their impact on decisions. Huang et al. [39] explain, "feature-sentiment information can help users digest user-generated reviews more efficiently". Neither former nor latter mentioned the impact of "sentiment scores" or "polarity detections" or evaluated their impact on decision-making.

Cambria et al. in their work [41] believed that sentiment analysis tools could be used to extract useful information from unstructured data. They were able to design a system that they believe is different "as it is an open-domain resource and it exploits reasoning techniques able to

infer general conceptual and effective information, which can be used for many different tasks such as opinion mining, affect recognition, text auto-categorization, etc." In their system they used a blend of common and common sense knowledge to build a comprehensive resource that can be seen as an attempt to emulate how tacit and explicit knowledge is organized in human mind and how this can be used to design an opinion mining and sentiment analysis system. This study also does not provide any evidence to support their claimed application- information extraction out of unstructured data- for their system.

Roses et al. [42] reviewed some applications of sentiment analysis. 1- Branding and product analysis: This application was studied by Hu and Liu [97]. They were able to build an algorithm using various techniques to summarize reviews on a specific product in terms of polarity. They believed that "summarizing the reviews is not only useful to common shoppers, but also crucial to product manufacturers" [97]. However, no empirical support for this claim was provided.

2- Analysis of political debates: In [98] Carvalhi et al. proposed a system that was mainly designed for analyzing political debates and they indicate that their system can be used by "the community interested in mining opinions targeting politicians from user generated content to predict future election outcomes" [98]. This application was tested later by various scholars such as Thumasjan et al. [108] and Thomas et al. [109].

3- Question answering: This topic was studied by Yu and Hatzivassiloglou [99]. They were able to design a system that could label documents in three main categories: fact, opinion and uncertain. Then the opinion sentences were evaluated in terms of positivity and negativity. Mix orientation, no orientation and uncertain orientation could also be labeled by their proposed system. They suggested that this system could be used by individuals who are looking for specific

answers to their questions out of unstructured massive answers available in question answering websites. This application was not tested empirically.

4- Tracking sentiment timelines in on-line forums and news: This topic was studied by Lloyd et al. [100], in which they suggested a model combining the reporter information (who, when, where) as well as sentiment detection on news website to predict or evaluate a social event. This system is tested and studied several times for event prediction purposes.

Paltoglu et al. [43] also believe that predictions can be made by summarizing the opinions that are available on social media interactions. They also believe that by extracting the polarity of one's opinion expressed through social media interactions such as tweets, my space comments and etc. without looking at the product review expressed by the same user, we are able to predict whether the user is going to recommend that product or not. The biggest challenge with working on social media interactions, such as tweets, is the level of unstructured communication. For instance, they refer to the work of Thelwall [13] that reports that 95% of the exchanged comments in MySpace contain at least one abbreviation (such as "m8" for "mate") of Standard English.

According to Grandos et al. [107] "it is evident that decision makers, i.e. consumers, suppliers and intermediaries, benefit from more information to make purchase and sales decisions". Libeman et al. [44] believe that there are three main challenges in collecting more information for decision makers to make a decision, and one stated by Engelberg [106] is "As most qualitative information is compiled in the shape of unstructured textual data, their processing is more costly as processing quantitative content". So Libeman et al. [44] decided to use sentiment analysis to summarize e-commerce unstructured qualitative data and transform it into quantitative data that was used by financial analysts and investors to make decisions. They were able to find that these results can be helpful in resource allocation on e-commerce. They also found that

individuals in different roles could interpret a unique piece of information differently to do financial predictions. However, the results of sentiment analysis were helpful for both groups.

Zhao et al. [45] believe that "tweets not only convey factual information, but also reflect the emotional states of the authors". They also believe that information about users' emotions is very important in understanding user behavior. In that, they refer to the work by Bollen et al. [101] in which the authors argued that "the events in the social, political and cultural fields did have a significant effect on the users' mood". Hence Zhao et al. claimed that those events could be predicted via users' underlying moods in their tweets. So they designed a sentiment analysis system that was able to categorize Chinese tweets into 4 levels of emotions: angry, disgusting, joyful and sad and they were able to find mood patterns in a time frame and detect the abnormal events according to those patterns. Again, no empirical evidence is provided to support the usefulness of these results for a specific task.

In his work, Liu [40] listed a set of application for sentiment analysis. From evaluation of consumer products, services, healthcare, and financial services to the analysis of social events and political elections all was mentioned and partially tested and reviewed by the author. He particularly believes that the information derived from sentiment analysis can be used mainly for predictions. He argues "such analyses can predict sales performance, volume of comments in political blogs or box-office success of movies as well as characterizing social relations". He did not test any of these claims using human subjects.

As described in this review section, studies on sentiment analysis mainly focus on system design and improvement of an existing system. Almost all studies claim at least one application for this technology. Yet, not so many supported their claim empirically. Hence, in this study we will use an off the shelf state-of-the-art sentiment analysis system to test the claim that sentiment

scores (the numerical result from sentiment analysis system) will improve decision-making. Specifically, we will evaluate whether sentiment scores are able to help individuals make a purchase decision faster and with more confidence. This evidence will mainly contribute to text analytics and sentiment analysis literature while providing support for further investment on system design and improvement studies. Also, the results from this study introduce new areas of research such as visualization of sentiment scores and how it impacts the decision making process.

3. Phase one- Star Ratings versus Sentiment Analysis - A Comparison of Explicit and Implicit Measures of Opinions

3.1. Introduction

In this phase of our study as stated in the problem specification and question section we are interested to test the validity and potential impact of sentiment analysis scores on information acquisition processes. Our measure for validity and potential impact is the correlation evaluation of sentiment scores on written documents with their respective star rating values. It is assumed that individuals' opinion regarding a product or service is substantially equivalent to their respective star ratings evaluation of that same product or service. Hence, sentiment scores on one comment, if valid, should be correlated with its respective star ratings value and may have the same impact for producing cues in information acquisition processes.

The hypothesis that we will test in this phase of our study is:

H1: Sentiment analysis scores on written reviews are correlated with their respective star ratings.

As sated, if the above hypothesis can be supported by the result of this study, then the result of the studies on usefulness of star ratings can be extended to sentiment analysis scores. For that purpose the simple statistical correlation between the two scores can be used to compare the two measures. The descriptive analysis will be presented later in section 4.3.

3.2. Methodology

To conduct this phase of our study, we used publically available archival data. The general guidelines we used in selecting reviews for products/services were that 1. there are abundant amount of reviews on the product/service 2. the purchase is not trivial for the consumer, and 3. the decision regarding the product/service has emotional as well as rational components.

3.2.1. Data selection process

We first selected four different products from the Amazon website (http://amazon.ca/) that had at least 40 reviews with corresponding star ratings. The first product was a pdf reader that had diverse reviews ranging from 1 star to 5 star ratings. The second product chosen was a book. Our prediction was that the reviews for this product would be slightly different from those of a technical product. This is because the reviews on books are sometimes regarding the storyline or the content of the book, which is occasionally different from the general opinion regarding how the reviewer enjoyed reading and consequently rated the book. Therefore, we expected to see different results from the sentiment analysis of this dataset compared to that of datasets about technology products. The third product studied was a streaming audio player, and the last one was an HDMI cable adaptor, two more technology products with a wide range of comments from positive to negative (with star ratings from 1 to 5).

The results of a study by Qiang et al. suggest that online user reviews have an important impact on online hotel-bookings [91]. Therefore, the second domain chosen for the analysis was hotel reviews. The data were gathered using tripadvisor website (http://www.tripadvisor.com/). Trip advisor is one of the best-known websites used by individuals to book hotels and get information regarding the destination they are going to visit. We selected three different hotels and collected on average 80 distinctive comments for each hotel to run the sentiment analysis. A general star rating regarding the consumer's overall experience in that hotel accompanied each comment. The three hotels were chosen randomly from a five star hotel to a 2 star one.

Lastly, we included reviews of doctors since the content of those reviews are different from those of hotels and products in that they are mostly about (albeit professional) person and thus

contain more sentiments than a typical consumer good. To gather data for doctors' review we used RateMDs website (http://www.ratemds.com/). RateMDs contains a database of doctors with different specialties and gives users who are supposedly the patients of those doctors the opportunity to rate and review them. Three family doctors, from Toronto, were randomly selected from this database. For each doctor, we collected 50 comments on average. The difference between doctors' reviews and hotels' reviews was that for doctors, there was not a general star ratings available but a rating was reported in four different categories: staff, punctuation, helpfulness, and knowledge. We used the (rounded) average of those ratings as the general star rating score to compare it with the comment's sentiment score.

3.2.2. Sentiment Analysis Tool

As indicated, in this study we do not aim to design or develop a new sentiment analysis tool but rather assess the available state of the art technology. Therefore we decided to use an off the shelf, publically available system, named Lexalytics (specifically, Lexalytics web demo[6]) as our sentiment analysis tool.

There are various open and commercial text mining and natural language processing tools that can perform sentiment analysis. The most commonly used tool in scholarly papers is Opinion finder.[7] This tool is mainly used to analyze tweets and is not able to analyze the text from our datasets that may sometimes exceed 20,000 words.

Some more examples for off the shelf sentiment analysis systems are Sentistrength[8] and sentiment 140[9]. Sentiment 140 is basically developed for tweets and is not able to analyze

[6] http://www.lexalytics.com/demo

[7] http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_1/

[8] http://sentistrength.wlv.ac.uk/

documents that contain more than 140 words. Sentistrength provides two separate scores for positivity (from 1 to 5) and negativity (from -5 to -1) while in our study we needed a unique score for the whole document. This drawback, along with the system's inability to work with longer than 140 word tweets makes this and the other above mentioned tools not qualified for our experiment. Lexalytics, on the other hand, delivers one single specific score in 3 decimal places between -1 to +1.

Besides these "pure" sentiment analysis tools, software solutions that perform various types of media analytics also provide sentiment analysis as one feature for analyzing social media. However, these tools are only able to search the media for a query and deliver the general trend of how people are talking about the specific key words in that query. Some of these tools also deliver the binary tagging for each comment or document, but none are able to deliver a specific numerical rating that goes beyond the binary evaluation, and hence these systems are also disqualified for our study. Sysomos[10], viralheat[11], lithium[12], Gravity[13] and Datasift[14] are some examples of such software.

Lexalytics includes a very large dictionary of sentiment bearing phrases in five different languages (English, French, Spanish, Portuguese, German) along with their relative sentiment scores. These scores are pre-determined by how frequently a given phrase occurs near a set of known good words (e.g. good, wonderful, spectacular) and a set of bad words (e.g. bad, horrible,

---

[9] http://www.sentiment140.com/

[10] http://www.sysomos.com/

[11] https://www.viralheat.com/

[12] http://www.lithium.com/

[13] http://www.gravity.com/

[14] http://datasift.com/

awful) [92]. This software identifies the emotive phrases within a document, scores these phrases (roughly -1 to +1), and then combines them to discern the overall sentiment of a sentence. This automatic sentiment scoring will score each sentence the same every time it is exposed to the system and is not affected by any human biases. Besides, its unique categorization engine, which requires no training, along with the ease of use of the system makes it uniquely appropriate for our study.

The first step in determining the tone of a document is to break the document into its basic parts of speech (POS tagging). POS tagging is a mature technology that identifies all the structural elements of a document or sentence, including verbs, nouns, adjectives, adverbs, etc. Lexalytics uses well-defined, well-understood techniques that generate extremely high accuracy for tagging the various Parts of Speech. Each query used on this system comes back with a hit count. These hit counts are combined using a mathematical operation called a "log odds ratio" to determine the score for a given phrase. Lexalytics uses an algorithm to combine the phrase scores in the document based on an operation called "lexical chaining" that supports the consistency and repeatability of the analysis [92].

To support our claim that Lexalytics is an adequate choice for our experiment, we compared the results from Lexalytics to those from another system that analyze and yields analytically equivalent results (the outcome for both is a specific polarity score rather than a binary classification). Lymbix[15] [78] is a sentiment analysis tool that is able to analyze documents that are longer than tweets, but still limits the number of words to 20,000. This drawback makes this system not to be our primary selected tool for this study. As stated, in our datasets of comments and reviews we sometimes deal with comments that are longer than 20,000 words. However, this

[15] http://www.lymbix.com/

system is able to provide numerical scores (from -10 to 10) rather than binary (positivity and negativity) reports.

Using this system, we were able to verify the results from Lexalytics by comparing it to the results from this new system. This comparison was made to test the accuracy of Lexalytics compared to another majorly used system. To conduct this comparison, we randomly selected one of our datasets and applied sentiment analysis to the data therein with both tools. Then a bivariate correlation test was conducted for the two sets of scores. The results are shown in table 2.

As is illustrated in the descriptive statistics box, Lymbix was not able to analyze 3 comments out of our sample of 88, because they contained more than 20,000 words. However, the correlation analysis shows that the results from the two systems are highly correlated, which further confirms that Lexalytics is a good representative of a sentiment analysis system that delivers specific sentiment score rather than a binary classification.

Table 2 - Lexalytics and Lymbix comparison

Descriptive Statistics

	Mean	Std. Deviation	N
Lexalytics	.346080	.2868738	88
Lymbix	2.817057	4.2070111	88

Correlations

		Lexalytics	Lymbix
Lexalytics	Pearson Correlation	1	.343**
	Sig. (2-tailed)		.001
	N	88	88
Lymbix	Pearson Correlation	.343**	1
	Sig. (2-tailed)	.001	
	N	88	88

**. Correlation is significant at the 0.01 level (2-tailed).

3.3. Analysis of data

We conducted sentiment analysis on each comment for each dataset using Lexalytics. Lexalytics provides sentiment scores in the range of -1 to +1. We normalized these scores to be expressed in the range of 1 to 5 (rounded the scores to their nearest integer values) to make them compatible with star ratings.

To compare each sentiment analysis score with its corresponding star ratings we conducted cross tabulation and chi square analyses for all the datasets. The chi-square test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. In our case, it specifically tests whether the sentiment analysis score (when normalized to a range from1to 5) for a comment is the same as the corresponding star ratings.

To test how much the results of sentiment analysis on the comments are related to respective star ratings, we also conducted two tail bivariate correlation analysis using SPSS. Bivariate correlation analysis compares the trends of the two datasets (sentiment analysis scores and star ratings).

Table 3 - Cross Tabulations of Normalized Sentiment Analysis Scores (SA) & Star Ratings (SR)

3.1. Hotel 1 Data Set

			SA				Total
			2	3	4	5	
SR	1	Count	2	2	1	0	5
		Expected Count	.1	1.7	2.8	.3	5.0
	2	Count	0	10	2	0	12
		Expected Count	.3	4.1	6.8	.8	12.0

	3	Count	1	15	2	0	18
		Expected Count	.5	6.2	10.1	1.2	18.0
	4	Count	0	8	25	2	35
		Expected Count	.9	12.1	19.7	2.4	35.0
	5	Count	0	6	37	6	49
		Expected Count	1.2	16.9	27.6	3.3	49.0
Total		Count	3	41	67	8	119
		Expected Count	3.0	41.0	67.0	8.0	119.0

3.2. Product 4 Data Set

			SA				Total
			2	3	4	5	
SR	1	Count	0	5	3	0	8
		Expected Count	.3	3.8	3.5	.3	8.0
	2	Count	0	7	2	0	9
		Expected Count	.4	4.3	3.9	.4	9.0
	3	Count	0	1	2	0	3
		Expected Count	.1	1.4	1.3	.1	3.0
	4	Count	0	5	5	0	10
		Expected Count	.4	4.8	4.3	.4	10.0
	5	Count	2	4	8	2	16
		Expected Count	.7	7.7	7.0	.7	16.0
Total		Count	2	22	20	2	46
		Expected Count	2.0	22.0	20.0	2.0	46.0

We included only two representative cross tabulations (one with significantly different distributions, and one without significantly different distributions) of sentiment analysis scores and star ratings in table 3. The results show that sentiment analysis results mostly fall into a neutral and moderately positive range of scores (3 and 4) rather than the extremes of 1 or 5. For most data sets, although there were many “1 star” ratings, there was a very low frequency of 1s on the corresponding sentiment analysis scores. The same trend is observable for scores of 5. As such,

distributions of the sentiment analysis scores seem fundamentally different from those of star ratings.

To test the statistical significance of that observation, we conducted chi-square tests, the results of which are displayed in Appendix 2. The results show that in 7 out of the 10 data sets that analyzed, the distribution of the star ratings and (normalized) sentiment analysis scores are significantly different from each other. Therefore for these data sets, sentiment analysis results seem to provide un-identical information to star ratings. The difference seems to be mostly due to a "neutralization" effect that sentiment analysis indicates. The next issue we address is, whether, in spite of these differences, the general tendencies (positivity and negativity) indicated in star ratings can be predicted by sentiment analysis.

For this purpose, bivariate correlation analyses were conducted. Table 4 displays the results. As seen in the table 4, for 9 of the 10 data sets studied, the sentiment analysis results are significantly correlated with star ratings ($p<0.01$). The only data set that yielded non-significant results belongs to a product where the reviews were shorter than those of the other products. Given that some of these reviews were less than 30 words long, they likely did not contain many sentimental phrases. That might be the reason that our sentiment analysis tool was not able to detect the sentiment of those comments.

The results indicate that although sentiment analysis results do not exactly correspond to opinions expressed in star ratings, these two scores are generally in agreement. For example, sentiment analysis of a review with a "1 star" rating almost always yields a negative score, although the degree of negativity is typically lower. Likewise, sentiment analysis of a review with a "5 star" rating almost always yields a positive score yet with a lower degree of positivity. In

other words, natural language expression of opinions seems to carry a more neutral tone even when an extreme star rating has been assigned to them.

Table 4 - SPSS bivariate correlation results

Domain	Sample size	Star rating Mean	Normalized SA Mean	Pearson Correlation
Dr-1	40	4.23	3.53	.415**
Dr-2	62	3.69	3.26	.620**
Dr-3	46	2.76	2.15	.442**
Hotel-1	119	3.93	3.67	.597**
Hotel-2	70	1.90	2.87	.640**
Hotel-3	65	4.58	3.85	.475**
Product-1	53	3.64	3.51	.523**
Product-2	48	3.19	3.31	.578**
Product-3	46	3.70	3.74	.585**
Product-4	46	3.37	3.48	.193

**. Correlation is significant at 0.01 level (2-tailed)

3.4. Discussion and Conclusions

The results show that the sentiment analysis has limited ability to detect extreme ratings explicitly assigned by reviewers. Meanwhile as reported in the background section, research indicates that those very extreme ratings are the most useful in helping consumers with their purchase decisions. Therefore current sentiment analysis is not a strong alternative to explicit consumer ratings, and should not be used to replace them.

One potential reason between the discrepancy between the explicit ratings and scores extracted from open-ended comments may be that people tend to use more neutral language while expressing their opinions in natural language. If that is the case, to be compatible with star ratings,

sentiment analysis techniques need to be more sensitive to the subtleties in natural language expressions. This, of course, is a significant challenge. Yet, if the idea is to use current technology to find surrogates for star ratings when they are not available, one simple solution would be to apply a simple linear or nonlinear filter to sentiment analysis results in a way to highlight the subtle differences away from the "neutral zone".

Another potential reason for the differences we observed may be stemming from the tool that was used in this study. To our knowledge, a comprehensive comparison of available sentiment analysis technology has yet to be performed. This part of our study suggests one criterion (ability to predict star ratings) that can be used in such a comparison. It is also possible that in order for any sentiment analysis tool to yield more meaningful results, the texts that are analyzed should be long enough to include a sufficient number of sentiment bearing phrases.

A related limitation of this study is that our data did not perfectly meet the distribution assumptions of chi-square test. This is largely due to the shortcomings of the sentiment analysis as discussed above. Future fine-tuning of sentiment analysis techniques might alleviate this issue hence improving the reliability of chi-square testing for comparisons such as what is reported in this part of our study.

In selecting our review data, we strived to choose domains where consumers typically use reviews. Nevertheless, our results should be generalized to other domains with caution. Future work should focus on developing theory that provides better guidance in selecting domains for empirical studies such as this one as the performance of sentiment analysis is likely to be domain specific.

Our results also imply that sentiment analysis is much better in capturing the general sentiments (negative, neutral or positive) expressed in star ratings. Therefore, sentiment analysis

scores for reviews without explicit ratings can be used in the same way as star ratings as a cue for which review might be the most useful to read. Sentiment analysis can also be used to assign a score to a part (for example each paragraph) of a long review hence detecting the variety of opinions within the same review. This helps consumers decide which part of a long review is more useful to focus on. Such fine-level support is not provided by current star ratings.

Sentiment analysis, as a "big data" analysis tool, holds much promise. In this study, we have attempted to explore the performance of current state of the art sentiment analysis technology in important domains where it can potentially be useful. We believe the importance of this technology will be more pronounced as user generated content gets bigger and more prevalent.

In the next phase of our study, we will empirically examine the impact of sentiment scores on purchase decision outcomes. More specifically, we will evaluate individuals' purchase decision outcomes when pages of reviews are used as their source of product information. In that, we will first conduct sentiment analysis on each paragraph of the review in the document file. Then we will apply the proposed partially- linear filter[16] to sentiment analysis results to highlight the subtle differences away from the "neutral zone" in order to produce a closer to star ratings cue for information acquisition. The purchase decision outcome will be investigated in terms of efficiency and effectiveness as will be explained in the following chapters.

[16] Extended score = Original score / 0.5 if original score < 0,

Original score/ 0.3, if original score >= 0

4. Phase 2 – User Studies

4.1. Introduction

This phase is designed to address the main question of this study, which is the evaluation of the usefulness of sentiment analysis technology. More specifically, we will investigate whether the outcomes of sentiment analysis technology – sentiment scores – are able to improve decision outcomes. The results from this phase of our study contribute to sentiment analysis and text analytics research by providing empirical evidence to support the usefulness of this technology in providing cues for information acquisition used for purchase decisions. Also, this study can be referred to as a justification for the investments (time and resources) on further studies in the field of sentiment analysis that yields to higher accuracy and improvements of this technology.

It has been thought traditionally that the main reason consumers search for new information is to reduce their uncertainty about their decision [5, 6]. This is due to the fact that sources of information are unlimited and it is almost impossible to get a thorough product or service information. Individuals, intuitively, find a balance between the cost of information acquisition and the benefit of each piece of information. Subsequently, they decide where, when, how, and to what extend they need to seek for information.

Individuals often use review websites as a good source of information to reduce the level of uncertainty in their purchase decisions [26]. However, the rapid increase in the volume of review information available as well as the substantial amount of user generated reviews, creates the problem of information overload. To mitigate this problem, consumers use decision and comparison aids [34] and numerical content ratings (such as star ratings) [35] not only to conserve cognitive resources and reduce energy expenditure to acquire information, but also to ease or improve the purchase decision process [26].

Sentiment analysis technology has been claimed to help users to make wiser decisions [39]. We believe that this can be due to the fact that sentiment evaluations on a piece of document can be used as a cue for information acquisition purposes and therefore for more effective and efficient decision outcome. In this study, we will assess this claim by investigating the impact of one particular sentiment analysis outcome-sentiment scores- on a specific decision process -purchase decision.

4.2. Theoretical background

Big data analytics mainly refers to two technical entities alongside each other. First, there's big data for massive amounts of detailed information and second, there's advanced analytics, which is a collection of different tools and quantitative techniques. The latter includes tools and technologies designed for predictive analytics, data mining, statistics, artificial intelligence, natural language processing, etc. Big data analytics has become the hottest new practice in Business Intelligence research [50]. However, the debate on the usefulness and success of those tools and techniques continues among scholars. Also there is always an argument on when, where and to whom these technologies can be more helpful and beneficial.

Big data challenges are not limited to specific group of individuals. We all may have experienced challenges caused by big data at least once in our modern life. One example is when we are making a purchase decision. The information gathered for our decision may come from different sources; from a friend's suggestion to the sales person's recommendation and most importantly product reviews and descriptions available online. Nowadays, online reviews and comments have become one of the most reliable and useful sources of product information. However, this huge amount of data can be more useful if clustered, organized and structured in the way that is more cognitive to our mind.

We, as consumers, sometimes may use decision and comparison aids and numerical content ratings (such as star ratings) [35] not only to conserve cognitive resources and reduce energy expenditure to acquire information, but also to ease or improve the purchase decision process [26]. On the other hand, we normally welcome any other new technologies that can contribute to overcome our big data challenges. Nevertheless, not all our efforts in learning and using new technologies turn out beneficial. This problem is more emphasized for innovators and early adopters, from technology adopters' lifecycle [51].

In this study, we will investigate the impact of sentiment scores (resulting from a sentiment analysis tool) on decision outcomes. In that, we will evaluate how sentiment scores will impact individuals' purchase decision outcomes when they use online reviews as the source of product information. The framework that we will use to support our hypothesis is derived from Cost/ Benefit framework, Information Foraging Theory, and Task Technology Fit. In the following sub-sections, we will briefly review these theories by providing explanations on their basic tenets and structures.

4.2.1. Cost/ benefit framework

Cost-benefit theory of decision strategy choice provides a conceptual foundation for studying human decision behavior. This framework asserts that individuals weigh benefits and costs before choosing a strategy for processing information in a decision making task [74, 75, 76]. In cost/benefit literature, "cost" refers to the mental effort by individuals for information acquisition and computation, while "benefit" can be considered as the impact of the strategy they choose to acquire the best and most useful information. In the context of cost/ benefit framework individuals choose a strategy or technology to acquire information that has more benefit and less cost cognitively.

Cognitive cost/ benefit perspective is based on the proposition that individual's performance will be affected if there is a fit between the strategy in cost/benefit and the task in hand. Moreover, Individuals can evaluate the fit between their task and a proposed technology, and are able to choose technologies or strategies of acquiring information according to that fit. In her work, Jarvenpaa [55] suggest that the cognitive cost/benefit framework can provide a robust theoretical foundation for design decisions regarding graphical presentation formats in decision support systems.

Cost-benefit theory also proposes a way to organize knowledge with regards to information presentation. Vessey denotes that certain strategies in problem-solving processes will dominate alternative ones when the problem representation matches the nature of the decision-making task [110]. Moreover, this theory suggests that decision makers may change their problem solving strategies to minimize the joint cost of effort and error in making a decision [110]. This is consistent with cognitive fit theory, which was explained in section 3.1 and can be considered as the economic justification for what naturally happens in human mind.

This theory has been widely used in technology and information system adoption and decision making studies. Cost benefit framework provides a formal description of the decision process and examines theoretical basis to understand how some decision aid tools have become more acceptable around the world over the others. Essentially, the aim of cost-benefit analysis is to provide a consistent procedure for evaluating decisions in terms of their consequences [111].

4.2.2. Information Foraging theory

"Information Foraging Theory is an approach to understand how strategies and technologies for information seeking, gathering, and assumption are adapted to the flux of information in the environment" [24]. This theory is originated from Optimal Foraging Theory or

more specifically Food Foraging Theory in anthropology [64] and behavioral ecology [65]. Food foraging theory provides a framework to explain food seeking and prey selection among animals. It evaluates the factors that influence the behavior of animals searching, selecting and consuming their foods.

The Optimal Foraging Theory indicates that animals will choose their food in the way that the amount of energy they gain from the food outweighs the amount of effort and energy they spend to search, select and consume the food [24]. This evaluation depends on the animal's body shape, habitat and the type of food itself [67]. Pirolli and Card [24] argue that information seeking in human mind is similar to food foraging behavior in animals and proposes Information Foraging Theory.

"The basic hypothesis of information foraging theory is that when feasible, natural information system evolves toward stable states that optimize gains of valuable information per unit cost" [24]. The theory assumes that individuals, when possible, will modify their strategies of acquiring information or the structure of environment to maximize their rate of gaining valuable information. Optimal information foraging focuses on how people will best shape themselves to their information environments, and how information environments can best be shaped to them, to get maximum amount of information in a limited amount of resource allocation (energy and time expenditure) [112].

This seems to be consistent with cost benefit framework discussed in section 3.2. As stated, individuals weigh benefits and costs before choosing a strategy for processing information in a decision making task. Human minds evaluate the amount of time and effort spent to obtain and process information as the cost of search for more information, and the value of the information

gained as the benefit. The result of the analysis conducted cognitively will result in finding an optimal point, which yields to minimum cost and maximum benefit.

Likewise, according to optimal information foraging, it is expected that in the process of searching, finding and digesting information, users need to adjust the strategy of information foraging to optimize the profit of information acquisition. Cognitive systems engage in information foraging are argued to be the reason to exhibit such adaptive tendencies [24].

Information Foraging Theory, as Pirolli and Card [24] explain, "attempts to specify the ways in which users search for information". They refer to the results from the study by Pirolli [68] that users are heavily influence by the "information scent". Pirolli [68] indicates that cues in the immediate environment of information presentation, will let out a scent about the nature of information. This scent then, will direct the user to either choose and pursue that source of information or ignore it for another more promising information paths to achieve information seeking goal.

Hyperlinked text on Web pages is an example for information cues that can possess various level of scent (a strong scent, weak scent, or no scent) based on the degree to which the hyper-linked words relates to the user's information goals [68]. Web browsing clustering is another means for information overload mitigation. Pirolli in another work [69] examined the impact of web-browsing clustering on information foraging and was able to support the claim that successful use of clustering may increase effectiveness and efficiency of information acquisition.

Sundar et al. [66] investigate the user reliance on information cues to moderate the information overload problem. More specifically, they evaluate the impact of different cues on news websites and how they affect users' information foraging behavior. The main problem in information gathering and sense making is the allocation of attention [24]. Different types of cues

have distinctive impact on users' attention. This variation depends on users' experience, information representation and the information goal. Sauder et al. [66] also, compared the impact of different cues on users' news selection behavior. They were able to find evidence that different combinations of cues have different effects on users' information selection behavior.

As indicated by Khapre and Basha [67] "Information clues, play a very important role in the process of directing the user to query information in the information foraging process". Individuals based on the understanding of the existing categories to their own minds (experience) and judgment of information available (mental representation of information), combined with specific tasks in different network environment may develop appropriate information feeding plan [67].

The analysis of information clues in Information foraging theory is based on four major theories [67]: The Lens Model by Brunswick [70], Anderson's classification of Adaptive Theory [71], Anderson's Memory Adaptation Theory [72] and Mcfadden's Random Utility Model [73]. Optimal information theory uses a combination of these theories along with some other foraging and behavioral models to articulate an estimation of how human minds behave in different information foraging situations.

4.2.3. Task Technology Fit

The ultimate argument of fit models states that Information Technology will be used and will provide benefits if the functions available to the user support the activities of the user [58]. The ability of Information Technology to support a specific task is expressed by the formal construct known as Task Technology Fit (TTF) [47]. This indicates that when technology characteristics match the task characteristics then there is a fit between that task and the technology and therefore using technology will improve users' performance.

Likewise, task-technology fit (TTF) theory holds that IT is more likely to have a positive impact on individual's performance and will be used in future, if the capabilities of the IT meets the requirements of the task that the user should perform. In another words, Task–Technology Fit (TTF) is defined as the degree to which a technology (broadly defined to include information technologies and information systems, but also other manual technologies or techniques used to assist in task accomplishment) assists an individual in performing his or her portfolio of tasks.

The positive relationship between task technology fit and user performance has been examined in many studies. Goodhue and Thompson [47] found support for this proposition for 25 different technologies in two organizations. Dickson et al. [60] studied the impact of information representation, comparing tabular versus graphical representation of a problem in 3 consecutive experiments. They were able to find evidence that supports the claim that no visual representation is superior to the other by its relative characteristics, but it is the match between these characteristics and the task in hand that makes them outperform the others. Moreover, they found that it is the fit between the problem representation and the task that improves performance not the type of representation alone. Staples and Seddon [61] in their study of the difference between the mandatory and voluntarily use of technologies and their impact of performance found support that if the task and technology matches, performance will be improved even if the use was mandatory.

The basic idea of task technology fit and the general model that other versions of TTF were built around is presented in figure 3.

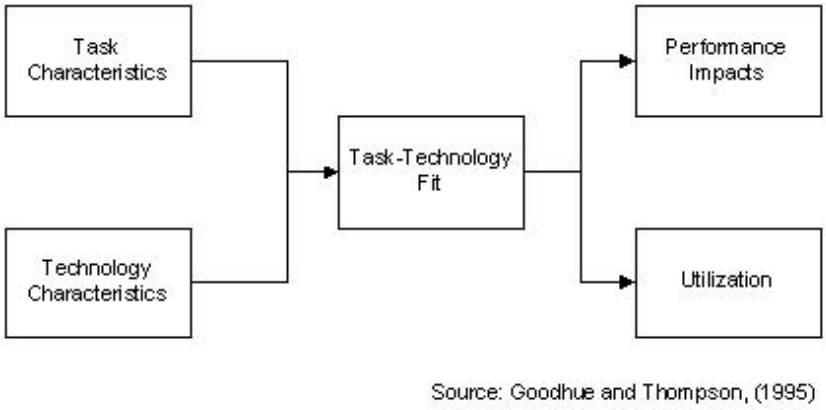


Figure 3 - Task Technology Fit

In this model, there are two factors that define task technology fit. One is technology characteristics and the other is task characteristics. As explained, if the technology features matches the task demands, then there is a fit between that technology and the task it is used to accomplish. Some other models of task technology fit have a common addition of individual characteristics to task and technology characteristics. The inclusion of individual characteristics is supported by Work Adjustment theory from which TTF was originally derived from [56].

The general concept of fit has been used widely by scholars in the field of management of information systems. Particularly, TTF has been used in the explanation of data representation techniques. For instance, Tan et al. [53] used the concept of cognitive fit and TTF framework to support their hypothesis on the dependency of data representation to task requirements. They specifically studied the use of graphs vs. tables for data representation and were able to provide support that the choice decision is dependent on the task requirements [53].

Task technology fit and cognitive fit theory, are commonly used interchangeably or as a support to each other such as in the study by Tan et al. [53] mentioned earlier. Both theories

provide foundation for the studies on the impact of the fit (between task and technology in TTF or problem and information in cognitive fit) on performance improvement for the former or efficient and effective problem solving in the latter. Vessey [54] was able to provide support that when data representation does not match the task requirements, it will slow down the decision making performance using both TTF and cognitive fit concepts interchangeably. This is consistent with the fact that, in the events when fit does not occur, users will need additional time to translate the presented data to useful information [54].

Goodhue indicates "information system, in general (systems, policies, IS staff, etc.), has positive impact on performance only when there is a correspondence between their functionality and the task requirements of users" [47]. He adds that "for an information technology to have a positive impact on individual performance, the technology: (1) must be utilized and (2) must be a good fit with the task it supports" [47].

Technology in this model refers to any tool used by individuals in carrying out their tasks. Some examples of technology are hardware, software and data as well as user's support such as training provided to assist users in performing their tasks [47]. Tasks are defined as actions carried out by individuals in order to turn inputs into outputs. Characteristics of individuals refer to specific abilities one may have, such as experience with the technology or experience with performing the task itself that will affect both individual's utilization of the technology and their performance.

As Goodhue indicates in his further research [52], the ability of the technology and its characteristics is not the only factor that should match the task requirements; he argues that user characteristics may also be another factor that influences the fit between the technology and the task in hand. This addition to the original TTF makes a more thorough model since user

characteristics such as experience and intelligence as well as demographic characteristics not only directly affect users' performance, but also have an effect on the perception of fit.

In our study of the impact of sentiment analysis technology on decision outcomes, we also believe that users' characteristics may affect decision outcomes; hence we will control for these characteristics as we perform our analysis.

4.3. Hypothesis and research model development

According to information foraging theory, individuals based on their experience and judgment of information available, combined with specific tasks that they are aiming to perform in different network environments may develop an appropriate information-feeding plan [67]. In addition, as Pirolli indicates, cues in the immediate environment of information presentation, will let out a scent about the nature of information [68]. This scent will then be used by individuals to design their information-gathering plan.

We believe that if sentiment scores can be used as a cue for information acquisition process, optimal information foraging will occur. Subsequently, optimal information foraging will provide the best and most beneficial (after a cost benefit analysis) amount of information, which yields to an efficient and effective decision outcome. Decision makers' characteristics such as experience and intelligence, as was mentioned in the extension of Task Technology Fit theory [52], will also affect the decision outcomes and will be controlled in our experiment. Hence, the proposition we will test in this part of the study is *Sentiment scores will improve consumers' purchase decision outcomes, while using reviews*.

In the first phase, we examined the potential impact of sentiment scores by comparing them to an existing implicit measure. As stated, previous studies found evidence that star ratings are effectively used as a cue for successful information acquisition, which support efficient and

effective decision outcomes. The results from the previous phase provide support for the potential impact of sentiment scores. It is argued that these scores for reviews without explicit ratings can be used in the same way as star ratings (as a cue for which review might be the most useful to read). It is also argued that sentiment analysis can be used to assign a score to a part (for example each paragraph) of a long review and hence detecting the variety of opinions within the same review. This helps consumers decide which part of a long review is more useful to focus on.

To provide empirical evidence for the above claim and test this potential application for sentiment scores, we conduct an experiment. The results will provide support for the effective use of sentiment scores in information acquisition and consequently will indicate the usefulness of sentiment analysis technology for this specific application. Essentially, purchase decision outcomes will be examined in terms of efficiency and effectiveness. Efficiency will be investigated by the *time* it takes consumers to make their decision and effectiveness will be measured by their *confidence* about their decision. Information provided to support this decision is pages of reviews annotated by sentiment scores for each paragraph.

"User characteristics" are also important factors that may affect the decision outcomes. For instance, decision maker's experience about the product may affect both the time to make decision and the level of decision confidence. Likewise, demographical characteristics might affect the decision outcomes both directly and through the effect on information foraging behavior. Also, individuals' speed of reading and interpretation of information may affect the efficiency of the decision outcomes. These characteristics are all expressed under the user characteristics that are used as control variables in our model. The model that we test is demonstrated in figure 4.

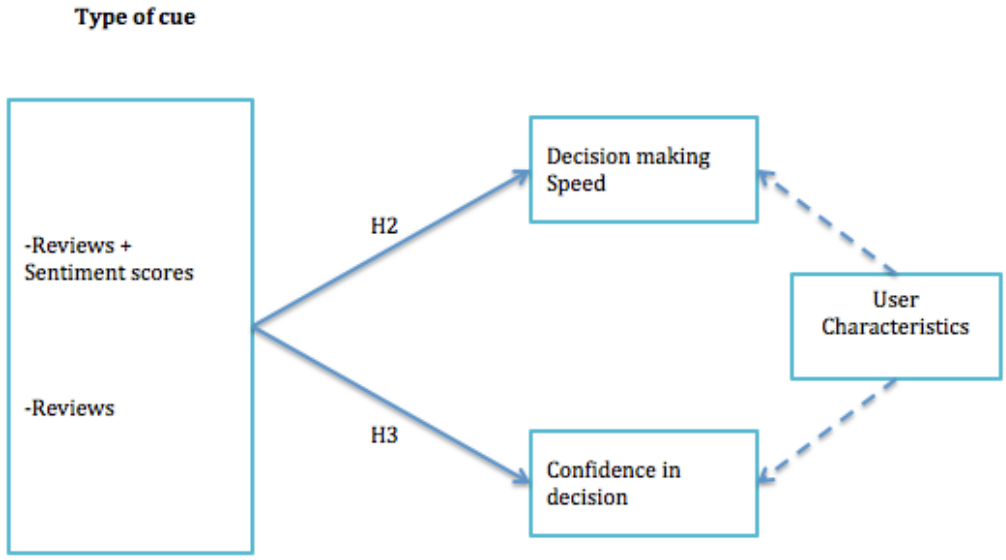


Figure 4 - Proposed Model

As illustrated in figure 4, two types of cues will be provided to decision makers and their decision outcomes will be investigated through two variables that are decision-making speed and confidence in decision. The former is an objective measure while the latter is a subjective measure evaluated and rated by users. The hypotheses that are tested in this study are:

H_2- Sentiment scores will help individuals make their purchase decision faster.

H_3- Sentiment scores will help individuals make their decisions with higher level of confidence.

In the next section, the research methodology will be explained and the experiment design as well as data gathering process will be outlined. Data analysis and results will be presented in the subsequent sections.

4.4. Research methodology

To explore the impact of sentiment scores, the main product of sentiment analysis technology, on decision outcomes, we conducted an experiment on random individuals making a conceptual purchase decision. The decision was on the purchase of a semi-professional camera and information provided was pages of product reviews as well as a thorough description of the specifications of the product. One major challenge with working on social media interactions, such as tweets or any other informal conversation, is the level of unstructured communication. Thelwall [13] reported that 95% of the exchanged comments on MySpace contain at least one abbreviation (such as "m8" for "mate") of Standard English. This finding and the level of inaccuracy that the structure of the text in those fields creates made us focus only on user and professional reviews rather than including any other social interactions such as tweets regarding the product.

Human subjects are involved in our study, since the information needed is regarding consumers' purchase decision outcomes. To estimate our sample size while providing support for our test of significance by providing reliable discrimination between Null (H_0) and the alternative hypothesis (H_1) of interest, we conducted a power analysis using G*power3 software[17]. By definition, the power of a statistical test is the probability that its null hypothesis (H_0) will be rejected given that it is in fact false [79]. "Statistics textbooks in the social and behavioral sciences typically stress the importance of power analyses" [79].

In a priori (analysis prior to the study) power analysis [80], sample size N is computed as a function of the required power level ($1 - \beta$), the pre specified significance level α, and the population effect size to be detected with probability ($1 - \beta$). In our study, we test the difference

[17] http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

between the decision outcomes of two independent groups; one using reviews annotated with sentiment scores for each paragraph while the other receives the review file only to support their decision. We defined the Cohen's effect size d measure[18] for our experiment to be 0.65, our α as 0.05, and 1- β (power) to be 0.95. d= 0.65 is considered as medium effect and this means that the treatment group outperformed the comparison group by 0.65 of a standard deviation. The result turns out to be a total sample size of 104 while groups are equally distributed and each contains 52 subjects. The result for our power test is provided in appendix-1.

A controlled experiment was designed for data gathering purposes. A questionnaire was prepared using questions mainly extracted from previous studies [88, 57, 87, 90, 89] and the constructs were defined according to the related sentiment analysis literature [39, 89]. Most of the variables were measured quantitatively using 7-point likert scale that is commonly used in previous IS literature. Decision outcomes were evaluated by one objective measure that is *time to make decision*, and a subjective measure, which is *confidence* and is measured by participants' evaluation of their decision.

In the previous phase of our study, we found that sentiment scores, while correlated with star ratings, are not able to predict the extreme sentiments. These scores are typically around the neutral values, and the extremes could not be detected. On the other hand, referring back to studies on the usefulness of star ratings, Pavlou and Dimoka found substantial evidence that the extreme ratings of eBay sellers were more influential and useful than moderate ratings [36]. The same result was identified in the study done by Forman et al. [37]. They specified that for books, moderate reviews (3 stars) were less helpful than extreme reviews. Although there is also evidence

[18] This measure can be calculated using the difference of two group's mean divided by total sample size. A lower Cohen's *d* indicates the necessity of larger sample sizes, and vice versa.

that neutral ratings are sometimes useful (such as in study by Crowley and Hoyer [38]), we decided to use "Extended sentiment scores" that amplify the negativity and positivity in text, in this second phase of our study.

A simple partially linear filter was applied to sentiment scores in a way to highlight the subtle differences away from the neutral zone. As a note from section 4.2.2, the original sentiment values are expressed as a number between -1, for extreme negative, and +1, for extreme positive sentiments. In our analysis on the documents used in this phase, which was consistent with the results from the previous phase, sentiment scores were more centralized around the neutral zone as shown in figure 5.

Thus, the extended sentiment scores used in this phase, were calculated as an extension of the original values (-1=< Original value =<+1) while the furthest from neutral (which is 0) is considered to be in one of the extreme (-1 if the furthest to neutral is negative and +1 if the furthest to neutral is positive) and the rest of the scores are expressed using the same scale to make the extremes. By that simple equation, we were able to add extreme rankings, which seem to be more influential than moderate ratings [36] to our experiment, while still benefiting from the original polarity classification. The contrast between the original and extended sentiment score is illustrated in the distribution graph shown in figure 5.

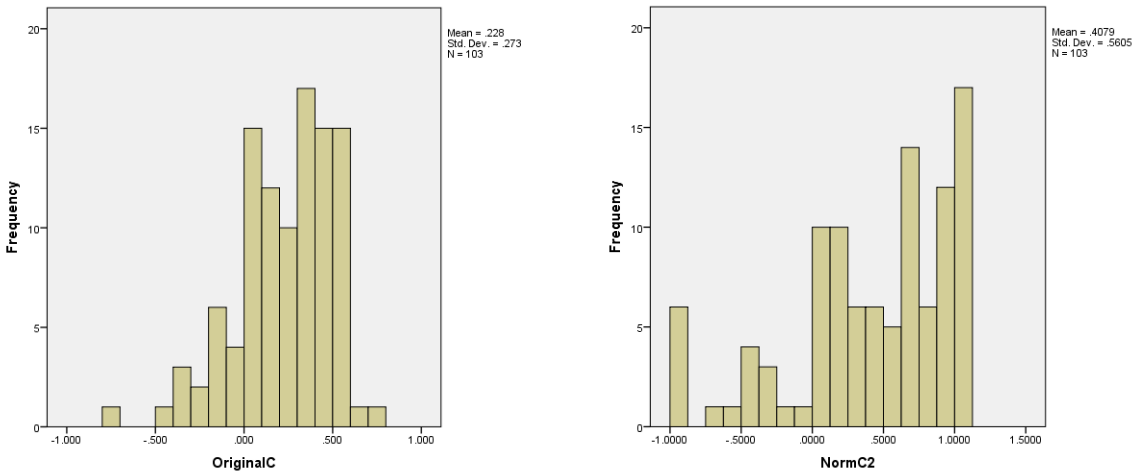


Figure 5 - Original vs. Extended sentiment scores

4.4.1. Experiment design

As stated, purchase decision outcomes are evaluated while pages of reviews are provided as the only source of product information. Online reviews are used by consumers in their purchase decision process for both utilitarian and hedonic purposes [81]. Individuals use online reviews since they assume other consumers' information is more important than advertising [83]. Eight distinct factors influence purchasers to search for more information online. Consumers seek the opinions of others online 1. to reduce their risk, 2. because others do it, 3. to secure lower prices, 4. to get information easily, 5. by accident (unplanned), 6. because it is cool, 7. because they are stimulated by off-line inputs such as TV, and 8. to get pre-purchase information [83].

"The most basic motive for a consumer to use reviews is the expectation of receiving information that may decrease decision time and effort and contribute to a more satisfying outcome" [82]. According to Schindler and Bikart [81] consumer reviews are the most frequently mentioned source of WOM that users seek for. They also mentioned that technical reviews would be used when the decision is important and risky.

In this experiment we selected semi-professional cameras and chose a moderate price range variations to have a product that might be considered both as an important product for some and "not so important" product for the others. In our review document, we used professional and technical reviews as well as user reviews along with a table of product specification. Three choices of competing cameras, from three different brands, were selected according to a camera review[19] website. The competition was not due to the cost and monetary value of the cameras, but the general capability of the products.

Product Specifications were extracted from dpreview (http://www.dpreview.com), one of the most trusted camera review websites. Professional Reviews were obtained from various camera review websites such as photo.net (http://photo.net), photographyblog (http://www.photographyblog.com/) and digitaltrens (http://www.digitaltrends.com). Finally, User Reviews were gathered from amazon website (http://www.amazon.com/). The final document was approximately 80 pages long with a hyperlinked directory that provided the readers with direct access to any part of the document.

As stated, we will test the difference between the decision outcomes of two independent groups; one using reviews annotated with sentiment scores for each paragraph while the other receives the review file only to support their decision. The review file presented to the former group was annotated by sentiment scores extracted from the same instrument used in the previous phase, Lexalytics. Two sets of questionnaires were created using Qualtrics[20] online tool and the prepared file was uploaded to the web based software environment. The data gathering process

[19] http://snapsort.com

[20] http://www.qualtrics.com/

was planned and application for Ethic Board's approval was prepared. Details will be outlined in the following subsections.

4.4.2. Measurement Scales

Two partially different questionnaires were generated according to literature and previous studies. We used Qualtrics web based survey service to design our questionnaires. Questions included in questionnaires are attached as appendix-4. User characteristics were measured by some background questions such as age range, gender, education level and past experience with DSLR (Digital Single Lens Reflect) cameras. Owning and intention to buy camera -in one-year timeframe- questions were used to evaluate the participant's interest as well as his/her experience in the prearranged purchase decision task.

The time was measured automatically by the online survey tool (Qualtrics). This was captured from the second subjects started reading the reviews to when they made their decision and got back to the questions. This time measurement tool was used in some other parts of the questionnaire to measure and monitor one of our control factors that is the natural reading and comprehension speed of the subjects. The second set of questions was regarding the subjects' decision and the information that was provided (the review document). The sufficiency, understandability and completeness of the information were measure for both groups. This was to compare the difference between the users' evaluation of the information provided within the two groups. These questions were extracted from the study done by Yang et al. [87]. We used 7-point likert scale to quantify these evaluations.

Decision makers' confidence questions were used to evaluate the effectiveness of decision outcomes. As stated, this was measured by users' evaluation on the same 7-point likert scale.

These set of questions were extracted from O'Connor's study on "Validation of a decisional conflict scale" [89].

4.4.3. Ethics approval

Ryerson University policies, requires that every research involving human participation, to receive Research Ethic Board's (REB) approval. The goal of REB is to protect the research subjects (participants) and to ensure that the research is conducted in an ethical manner (Research Ethic Board 2013). For that purpose a web based application form should be completed and submitted to REB. The committee asked for a meeting to revise some of the recruitment materials as well as description of the study. The required revision was applied and submitted back. The second revision was requested from the REB and all required materials were submitted for the third time. The final approval from REB was received after 45 days from the first submission of application. This approval allowed us to start the recruitment process.

4.4.4. Pilot study

To test our measurement scales and experimental design process, we conducted a pilot study. 8 graduate students with different backgrounds from engineering to business and computer science participated in this experiment in the role of both subjects and advisors for the general process and more specifically on the contents of the material and measurement scales.

After this pilot, the time for the sessions was reduced from one hour to 30- 45 minutes and the review file was changed from an Html version to a word document file, which was reported easier to read by our participants. Another change from the original data gathering process was that we decided to start the study for each participant separately, whenever they arrive to the venue, rather than wait for everyone to give a single instruction for the whole team. To make the situation consistent for all participants the instructions were given from a written document. Also a

control number was given out as raffle tickets pinned to the subjects' consent form rather than a number written on their forms to protect anonymity of participants.

Another feedback was regarding the amount of information provided. Almost all participants suggested that the review file was too long and would not allow users to get enough information in a limited amount of time and due to boredom. We consider this a positive one since the amount of information was too much by design given the nature of our study. In fact, behavior of users was monitored while they were presented with an overload of information. Hence, we decided to keep the amount of information as it was originally designed.

4.4.5. Data collection

The recruitment process started once the ethic's approval was received and the pilot was completed. An event page was designed using a web-based event planning software (www.eventbrite.ca). Several sessions were planned to match interested participants' schedule. An incentive was also arranged to appreciate and encourage participations. The incentive was set to be $10 cash and entering on a draw for an apple iPad. Advertisement posters – approved by REB- were attached on the news boards of almost all Ryerson departments. Invitation emails were sent to all student unions and professors to help in the recruiting process. A recruitment pitch was given during some summer classes upon professors' permission.

Signs were displayed across the campus to announce when sessions were taking place. As a result, in the first round of data gatherings, conducting more than 25 data gathering sessions, we were able to attracted 76 participants in total that were equally distributed in our two groups. The second round of data gatherings took place during fall semester. Same recruiting speeches were given and posters and signs were displayed. 5 more sessions were coordinated and 41 more data

points were added to the previous data, making the total 117, which is slightly greater than 104, the result from our power analysis.

All sessions were held at computer labs located at Ted Rogers School of Management. The whole purpose was to have a controlled experiment so none of the participants had the access to any other information rather than the review file provided by the investigator. Participants were required to read the consent form (attached in appendix-3) carefully, sign and date it to follow the ethics policies. Verbal instruction was also provided for each individual and the opportunity to ask clarification questions at any time during the study were given verbally as well as written through consent form.

Once the decision was made and final set of questions was answered, survey completion was controlled, consent forms were collected, and the incentive was provided. The validity of data points was investigated right after the completion of each session. Data points in which participants spent less than 1 minute on the review file were dropped instantly from the data file. Also, incomplete questionnaires and/or missing values for highly involved variables were detected and if substitution was not applicable the whole data were deleted. Consequently, we ended up with 100 usable data points, 50 from the group using sentiment scores and 50 from the other.

4.5. Data analysis and results

All statistical analyses were done using SPSS software. In the next sub-section, we will start by defining the coding of variables that are used in our analysis. This will be followed by descriptive statistics analysis of those variables and distribution of our data. In the final section, test of our hypotheses and statistical evidence is presented.

4.5.1 Coding of variables

As stated, in our experiment, we had two groups; one using sentiment scores along with the reviews while the other used reviews only. Group 0 refers to the group using sentiment scores while group 1 signifies the one using reviews without annotations. Participants were asked about their gender and education level (demographical characteristics) as well as their familiarity with DSLR cameras that was the subject of our study. In terms of gender, "male" was represented as group 1 and "female" was denoted as group 2. Education had 7 different levels from 1 as lowest being "some high school" to 7, highest level of education and being "postgraduate degree". Familiarity was measured by a subjective measure "Familiar" that was rated out of a scale of 7, 1 being strongly unfamiliar and 7 being strongly familiar. Table 5 lists these variables with their respective coding and descriptions.

Time to make decision was measured by a time recorder instrument that was built in to our survey environment. This timer started recording once users opened the review file and it stopped at the time they made their decision. This was detected when they got back to the survey environment and a clicked on "I made my decision" button. Same timer was used several times during the survey to control for users' "reading speed", which was the measure of both speed of reading and interpretation of the information that will be used as a control factor in our analysis. This variable is labeled as "timefirst" in our analysis.

Confidence was measured by 3 variables, using questions extracted from the study by O'Connor [89]. These variables are measured by user evaluation measurement technique and are evaluated out of scale of 7. The result from factor analysis shows that all three variables (labeled as conf1, conf2, conf3 as outlined in Appendix 4) loading on one main component denoted as confidence (table 6).

Table 5 - Description of Variables and Coding

Variable	Coding	Description
Group	0	Using sentiment scores
	1	Not using sentiment scores
Gender	1	Male
	2	Female
Education	1	Some high school
	2	High school graduate
	3	Some college
	4	Trade/ Technical/ Vocational training
	5	College graduate
	6	Some post graduate work
	7	Postgraduate degree
Familiarity	1	Strongly Unfamiliar
	2	
	3	
	4	Neither familiar nor unfamiliar
	5	
	6	
	7	Strongly familiar
Reading speed (timefirst)	Measured in seconds	N/A
Time to make decision (timefile)	Measured in seconds	N/A

Table 6 - Factor analysis on Confidence factors

Communalities

	Initial	Extraction
conf1	1.000	.706
conf2	1.000	.815
conf3	1.000	.603

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.124	70.802	70.802	2.124	70.802	70.802
2	.586	19.524	90.325			
3	.290	9.675	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component
	1
conf1	.840
conf2	.903
conf3	-.777

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

We then used SPSS to calculate that unique confidence variable using factor scores extracted from the factor analysis [86]. This tool will produce a new factor with the mean being 0 and the effect of each construct being multiplied by their factor scores. In fact, the system uses a linear mathematical equation to build this new factor from the three variables that are explaining it [63, 77]. The result turned out a confidence variable with distribution shown in table 7. The minimum shows the lowest level of confidence while the maximum expresses the highest level of confidence. This measure will be used as the subjective measure of decision outcome in our analysis.

Table 7 - Descriptive statistics for Confidence

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Confidence	100	4.09257	-2.70772	1.38485	.0000000	1.00000000
Valid N (listwise)	100					

4.5.2. Descriptive statistics and frequencies of variables

In this section, each variable is evaluated through a descriptive statistical analysis and frequency of data is provided. Descriptive statistics provides information such as range, variance, minimum, maximum and mean as well as standard deviation for each variable. Table 8 provides a summary of our variables' descriptive statistics.

Table 8 - Descriptive Statistics for all Variables

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Group	100	1	0	1	.50	.503
Gender	100	1	1	2	1.39	.490
Education Level	100	5	2	7	4.49	2.028
Familiarity with subject	100	6	1	7	3.88	1.950
Reading Speed	100	218.378	1.451	219.829	51.01799	36.888424
Time to make Decision	100	2881.931	63.200	2945.131	1097.95239	785.486160
Confidence	100	4.09257	-2.70772	1.38485	.0000000	1.00000000
Valid N (listwise)	100					

As illustrated in table 8, the variances for both time variables are too large. For instance, for "Time to make decision" the minimum is 63 seconds (1 minute) while maximum being 2945 seconds (49 minutes) that was the maximum amount of time provided to our subjects to make their decision. It is predicted that these two variables may not generate the expected results to support our hypotheses. This is according to the wide distribution of time that might be due to couple of factors such as experiment design.

Table 9, shows the frequencies for each variable. As shown, the distribution for group is equivalent and each group contains 50 participants, which is ideal for our tests. 61% of our participants were male while 39% were female. While most of our participants claimed a moderate

knowledge about the subject, DSLR cameras, 17% reported absolutely no familiarity and 6% reported strong familiarity with the subject. In our analysis, we will consider familiarity level lower than 2 as unfamiliar and familiarity of higher than 5 as being familiar. It is interesting to highlight that 30% of our participants were high school graduates while 26% were having a post graduate degree. Therefore, we can claim that our sample covered a wide range of variation from education point of view and that makes a great random sample from that perspective.

Table 9 - Frequencies of Variables

Group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Using SA	50	50.0	50.0	50.0
	Not Using SA	50	50.0	50.0	100.0
	Total	100	100.0	100.0	

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	61	61.0	61.0	61.0
	Female	39	39.0	39.0	100.0
	Total	100	100.0	100.0	

Familiarity with subject

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not familiar at all	17	17.0	17.0	17.0
	2	15	15.0	15.0	32.0
	3	10	10.0	10.0	42.0
	4	11	11.0	11.0	53.0
	5	21	21.0	21.0	74.0
	6	20	20.0	20.0	94.0
	Stronly familiar	6	6.0	6.0	100.0
	Total	100	100.0	100.0	

Education Level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	30	30.0	30.0	30.0
	3	12	12.0	12.0	42.0
	5	21	21.0	21.0	63.0
	6	11	11.0	11.0	74.0
	Highest level of education	26	26.0	26.0	100.0
	Total	100	100.0	100.0	

"Time to make decision" and "Reading speed" being scale variables, as discussed, have a wide range of distribution and hence the frequencies for those variable is specified by a scatter plot illustrated in figure 6.

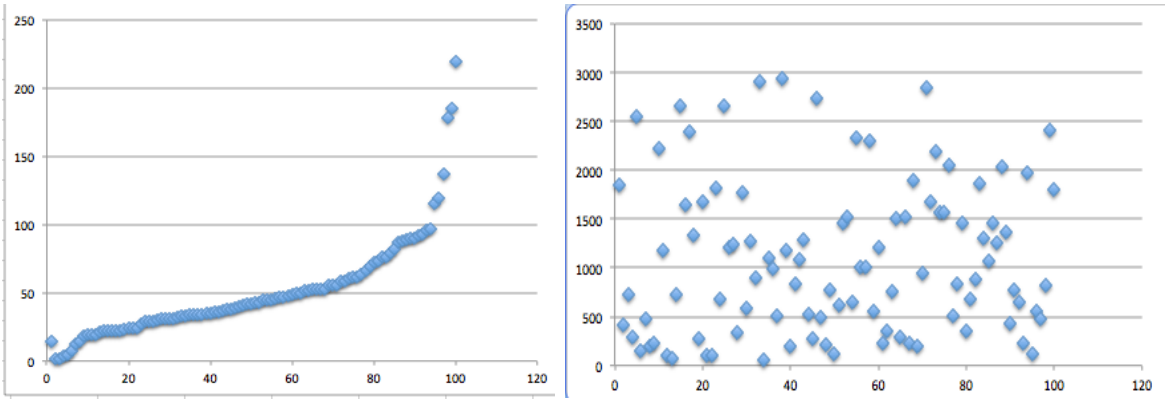


Figure 6 - Distribution of Reading speed (left) and Time to make decision Variable (right)

To test whether the two measures of decision outcome are different, we conducted a correlation test between these two measures. The result is provided in table 10. The scatter plot for this distribution is also provided in figure 7. The outcome from this evaluation provides support that these two measures are not significantly correlated and should be evaluated separately, since each represents a unique characteristic of decision outcome.

Table 10 - Relationship between "Time to make decision" and "Confidence"

Correlations

		Time to make Decision	Confidence
Time to make Decision	Pearson Correlation	1	.000
	Sig. (2-tailed)		.999
	N	100	100
Confidence	Pearson Correlation	.000	1
	Sig. (2-tailed)	.999	
	N	100	100

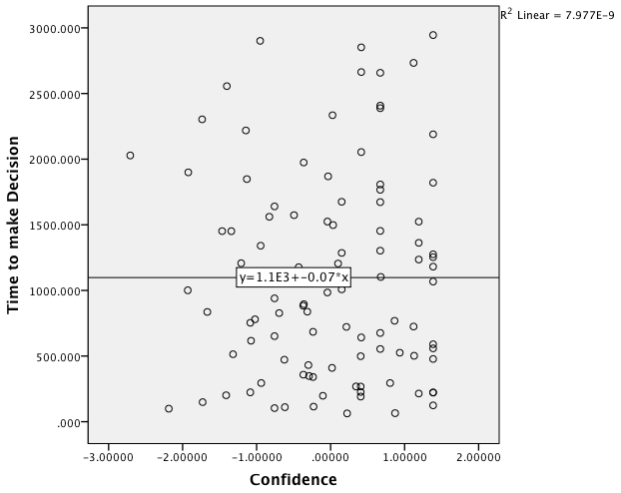


Figure 7 – "time to make decision" versus "confidence"

Moreover, participants provided feedback on which part of the review file was more useful in their decision making process. As stated, three main sections were provided for each camera, Professionals' review (prorev), Users' review (userrev) and product's General features (genfeat). The distribution of evaluations is presented in table 11.

Table 11 - Distribution of use of sources of information

Prorev

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	52	52.0	52.0	52.0
	1	48	48.0	48.0	100.0
	Total	100	100.0	100.0	

Userrev

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	31	31.0	31.0	31.0
	1	69	69.0	69.0	100.0
	Total	100	100.0	100.0	

Genfeat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	33	33.0	33.0	33.0
	1	67	67.0	67.0	100.0
	Total	100	100.0	100.0	

This table shows that 48% of our participants indicated that they used professionals' reviews in their decision process while 69% used users' reviews and 67% general specifications. It should be noted that each user was able to choose more than one source of information. This evaluation shows that reviews were considered useful by almost 70% of our participants and hence provides evidence that they used reviews to make decisions, and it is appropriate to assess the impact of reviews on their decision.

4.5.3. Hypotheses Testing

To test our main hypotheses in this study, we should investigate the difference between decision outcomes within our groups from two perspectives; time to make decision and confidence level considering the impact of our control factors on these two components. This requires a statistical test with the ability to compare means of our dependent variables (time to make decision and confidence level) within our two groups (using sentiment scores vs. not using sentiment scores) while controlling for the effect of other factors that are not of primary interest but will have

an effect on our dependent variables (such as familiarity, gender, education and reading speed). One statistical analysis test with the promise of supporting the requirements of our investigation is ANCOVA [84].

Analysis of covariance (ANCOVA) is a general linear model, which is a combination of Analysis of Variance (ANOVA) and regression. This statistical test evaluates whether population means of a dependent variable (DV) are equal across levels of a categorical independent variable (IV), while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates (CV).

Test of Covariance analysis (ANCOVA) has some assumptions that should be taken into consideration while using this test. One is the test for normality of dependent factors under analysis. The result for normality test on our dependent variables using "Q-Q plot normality test" is presented in figure 8.

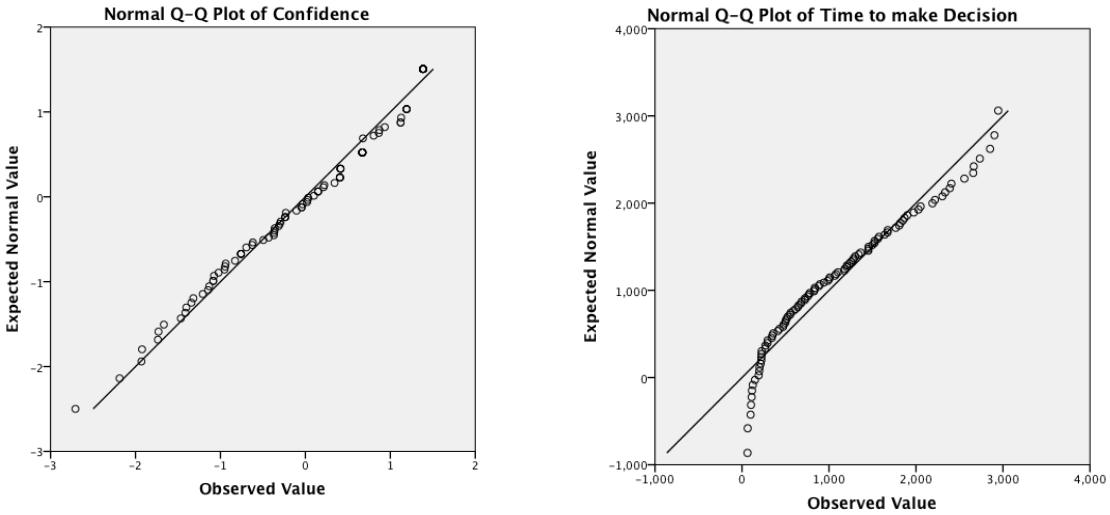


Figure 8 - Test for normality using Q-Q plot

According to figure 8, confidence (DV_1), follows the normal line and its distribution can be considered as being normal. Time to make decision (DV_2), however, slightly deviates from

normality line. This deviation is more detected around the lower levels and might have a trivial effect on our analysis since it might slightly increase type1 error. Hence, it is anticipated that if the seven observations (extracted from figure7 right) reporting the lowest time to make decision are dropped, the distribution will be closer to normal. However this yields to the reduction of our sample size and accordingly the increase in type 1 error in that sense. Hence, we decided to keep those observations and accept the increase in type 1 error resulting from not satisfying this assumption of ANCOVA.

Second assumption is the independence of the covariates and treatment effects. This can be tested through a t-test for our covariates between the two groups. This is to provide support for the random distribution of our covariates between groups. The result for that t-test is illustrated in table 12.

Table 12 - Test for the independency of covariates and treatment effect

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Reading Speed	Not Using SA	50	48.26048	42.927548	6.070872
	Using SA	50	53.77546	29.849508	4.221358
Familiarity with subject	Not Using SA	50	3.68	1.867	.264
	Using SA	50	4.08	2.029	.287

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Reading Speed	Equal variances assumed	.274	.602	-.746	98	.458	-5.514980	7.394278	-20.188684	9.158724
	Equal variances not assumed			-.746	87.405	.458	-5.514980	7.394278	-20.210948	9.180988
Familiarity with subject	Equal variances assumed	.325	.570	-1.026	98	.308	-.400	.390	-1.174	.374
	Equal variances not assumed			-1.026	97.335	.308	-.400	.390	-1.174	.374

As shown, both continuous control variables are non-significant, and therefore this assumption for ANCOVA is also satisfied by our control variables that are Familiarity to the subject and Reading speed. There are also other categorical and ordinal factors that might have an effect on decision outcomes such as Gender and education. These two will also be used as random factor in our analysis.

The third assumption for ANCOVA is "homogeneity of regression slope" for covariates. Figure 9 contains scatter plots and regression lines for each control variable in terms of each dependent variable. Regression lines for IVs (independent variable) are also graphed. If the slopes are equal (the lines are parallel) then this assumption is satisfied.

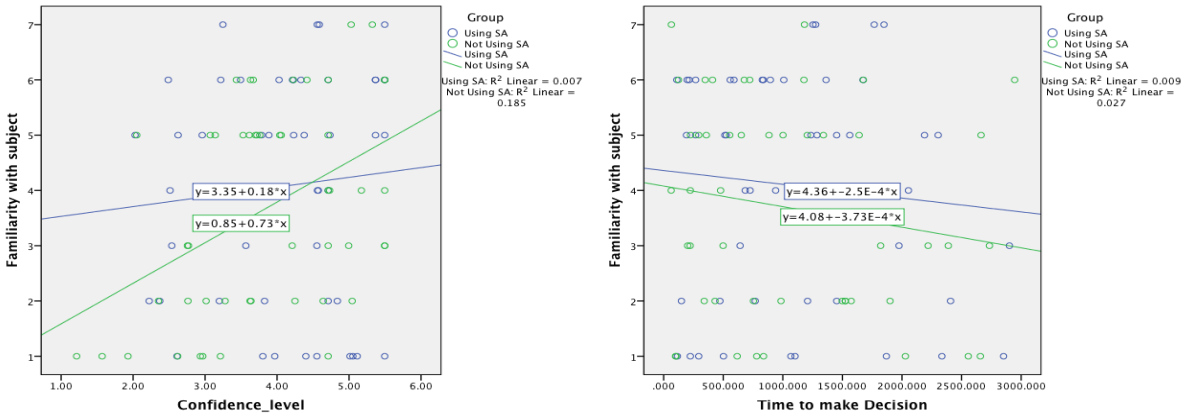


Figure 9.1 – Left: Familiarity and confidence level, Right: Familiarity and Time to make decision

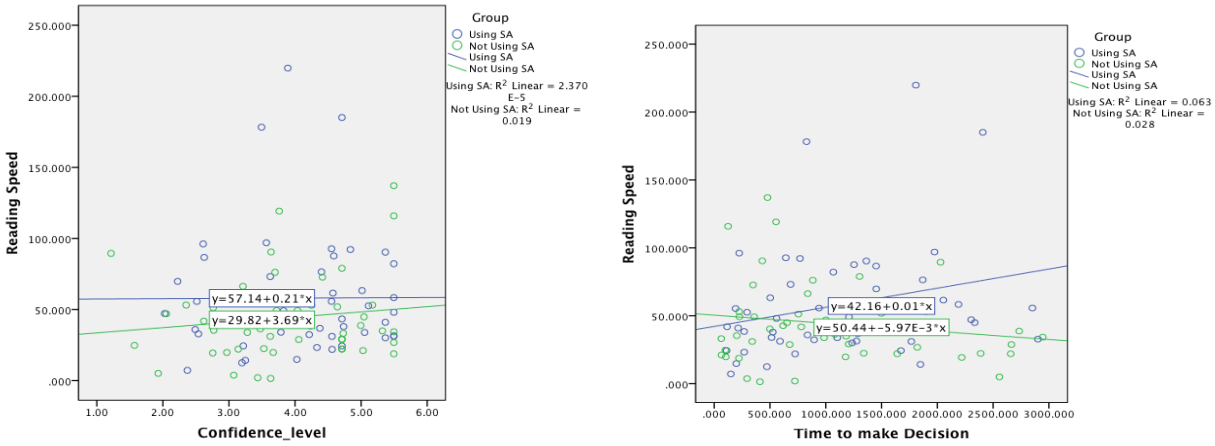


Figure 9.2 – Left: Reading speed and confidence level, Right: reading speed and Time to make decision

Figure 9 - Test for homogeneity of regression slopes

As illustrated in figure 9.1 left, the regression lines for two groups are not parallel and the slopes are not equal for familiarity while measuring confidence. This difference translates into the

heterogeneity of the variances and will lead to a greater error type 1. When such an incident occurs it is suggested to use an alternative analysis approach [85]. This seems to be the most important assumption of ANCOVA. As a result, we produced another measure for familiarity that is a categorical variable and can be used as a random factor in our analysis using ANCOVA. Also figure 9.2 right shows that reading speed cannot be used for the analysis of time to make decision for the same reason as discussed before. So this measure will not be used in our ANCOVA analysis for time to make decision.

The forth assumption of ANCOVA can also be tested through the same graphs. The assumption indicates, "The regression relationship between the dependent variable and concomitant variables must be linear". The scatter plots plus regression lines confirm the linearity of the regression relationship between DVs and Covariates. Lastly the fifth assumption is "the independence of error terms". This specifies that the correlation of error terms should be significant. The error term is independent of the covariates and the categorical independents. Randomization in experimental designs assures this assumption will be met that is the case for our experiment (Table 13).

Table 13 - Independent sample t-test on control variables

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reading Speed	Equal variances assumed	.274	.602	-.746	98	.458	-5.514980	7.394278	-20.188684	9.158724
	Equal variances not assumed			-.746	87.405	.458	-5.514980	7.394278	-20.210948	9.180988
Familiarity with subject	Equal variances assumed	.325	.570	-1.026	98	.308	-.400	.390	-1.174	.374
	Equal variances not assumed			-1.026	97.335	.308	-.400	.390	-1.174	.374
Education Level	Equal variances assumed	.078	.780	-1.614	98	.110	-.660	.409	-1.471	.151
	Equal variances not assumed			-1.614	97.524	.110	-.660	.409	-1.471	.151
Gender	Equal variances assumed	.164	.687	-.203	98	.840	-.020	.099	-.216	.176
	Equal variances not assumed			-.203	97.992	.840	-.020	.099	-.216	.176
Familiarity-group	Equal variances assumed	.268	.606	-.794	98	.429	-.14000	.17643	-.49011	.21011
	Equal variances not assumed			-.794	97.986	.429	-.14000	.17643	-.49011	.21011

After evaluation of the assumptions of ANCOVA, we can use it to test our hypotheses, it is time to apply the test considering limitations mentioned above. As a note from section 5.1, the alternative hypothesis (H_2) that we will test in this part is:

H_2- Sentiment scores will help individuals make their purchase decision faster.

According to this hypothesis we expect that participants from group 0 (using sentiment scores) will be significantly faster than those of group 1 (not using sentiment scores). In that, time to make decision for the former group should be significantly lower than that of group 1.

Table 14 illustrates the results from ANCOVA test on "time to make decision" factor, considering all control variables satisfying the ANCOVA test assumptions for this factor.

Table 14 - ANCOVA test for H_2

Tests of Between-Subjects Effects

Dependent Variable: Time to make Decision

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	27762057.5	1	27762057.5	21.071	.061
	Error	2211880.91	1.679	1317542.5[a]		
Group * Education	Hypothesis	3179352.28	4	794838.070	1.485	.214
	Error	46026087.8	86	535187.07[b]		
Group * Gender	Hypothesis	1197427.95	1	1197427.95	2.237	.138
	Error	46026087.8	86	535187.07[b]		
Gender * Familiar	Hypothesis	871436.895	1	871436.895	1.628	.205
	Error	46026087.8	86	535187.07[b]		
Group	Hypothesis	43072.229	1	43072.229	.033	.877
	Error	2007783.47	1.527	1315271.2[c]		
Education	Hypothesis	8190258.44	4	2047564.61	2.595	.185
	Error	3255281.00	4.126	788939.45[d]		
Gender	Hypothesis	1067446.68	1	1067446.68	1.610	.240
	Error	5252902.48	7.925	662830.54[e]		
Familiar	Hypothesis	592673.278	1	592673.278	1.107	.296
	Error	46026087.8	86	535187.07[b]		

a. .188 MS(Education) + .935 MS(Gender) - .123 MS(Error)

b. MS(Error)

c. .858 MS(Group * Education) + .841 MS(Group * Gender) - .700 MS(Error)

d. .977 MS(Group * Education) + .023 MS(Error)

e. .193 MS(Group * Gender) + .807 MS(Error)

According to this table, there is not any significant difference for "time to make decision" between groups. Therefore, H_2 is not supported. This is consistent with our prediction from the

distribution of this time construct and due to the limitations of our study that will be highlighted in limitations section (section 9).

The result from the test for H_2 did not provide evidence to support the hypothesis that sentiment scores helps users to make decision faster. One more decision outcome character, confidence, should be tested to find evidence that sentiment scores were successfully used as a cue for information acquisition purposes and hence can improve decision outcome:

H_3- Sentiment scores will help individuals make their decisions with higher level of confidence

This is due to our prediction that sentiment scores will be used as a clue to read through that huge review file and accordingly participants might feel more confident while having access to those review comparison aids.

This is also tested using ANCOVA while considering its assumptions for this specific variable. As mentioned, we will use a categorical familiarity variable for this test to avoid the problem of violating one of the assumptions of analysis of covariance test. It is important to note that the hypothesis that will be tested by ANCOVA is only able to detect whether there is a difference between the groups in terms of confidence. Further analysis should be implemented to find the exact effect. The alternative hypothesis tested by ANCOVA will be referred to as $H_{3\text{-}1}$ that is: $H_{3\text{-}1}$- Sentiment scores will help individuals make their decisions with *different* level of confidence

To test this hypothesis, dependent variable- confidence- will be tested over the independent variable- group- considering the effect of all control variables, i.e. speed of reading as a covariate and other categorical variables- education, gender and the new familiarity variable- as random factors. The result for this test is reported in table 15.

Table 15- ANCOVA to test H_{3-1}

Tests of Between-Subjects Effects

Dependent Variable: Confidence

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	.015	1	.015	.029	.906
	Error	.295	.570	.518[a]		
Group	Hypothesis	1.460	1	1.460	.495	.561
	Error	5.304	1.800	2.947[b]		
Education	Hypothesis	4.158	4	1.039	.387	.810
	Error	12.674	4.721	2.684[c]		
Gender	Hypothesis	.606	1	.606	.111	.766
	Error	13.181	2.419	5.450[d]		
new_familiarity	Hypothesis	3.627	2	1.814	1.203	.449
	Error	3.147	2.088	1.507[e]		
Timefirst	Hypothesis	.367	1	.367	.464	.498
	Error	54.479	69	.790[f]		
Education * Gender	Hypothesis	11.306	4	2.827	3.580	.010
	Error	54.479	69	.790[f]		
Group * Education	Hypothesis	4.694	4	1.173	1.486	.216
	Error	54.479	69	.790[f]		
Group * Gender	Hypothesis	4.092	1	4.092	5.183	.026
	Error	54.479	69	.790[f]		
Gender * new_familiarity	Hypothesis	2.569	2	1.285	1.627	.204
	Error	54.479	69	.790[f]		
Group * new_familiarity	Hypothesis	2.390	2	1.195	1.513	.227
	Error	54.479	69	.790[f]		
Education * new_familiarity	Hypothesis	6.133	8	.767	.971	.466
	Error	54.479	69	.790[f]		

As shown in table 15, the interaction between group and gender yield a significant difference in confidence level. Hence, the hypothesis is supported for some cases depending on gender. Further analysis should be performed to find the specific nature of this result.

To further investigate how sentiment scores influences each gender group, we conducted the analysis two more times, one for each gender group individually evaluating the main effect of each variable on confidence. The result from ANCOVA on Gender 1, males, is reported in table 16.

Table 16- ANCOVA to test H_3 on Males only

Tests of Between-Subjects Effects

Dependent Variable: Confidence

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	.390	1	.390	.157	.707
	Error	14.340	5.752	2.493[a]		
Group	Hypothesis	.437	1	.437	.518	.475
	Error	43.887	52	.844[b]		
Education	Hypothesis	11.197	4	2.799	3.317	.017
	Error	43.887	52	.844[b]		
new_familiarity	Hypothesis	9.788	2	4.894	5.799	.005
	Error	43.887	52	.844[b]		
Timefirst	Hypothesis	.515	1	.515	.610	.438
	Error	43.887	52	.844[b]		

As shown in table 16, group is not significant for the analysis on male participants. Hence, the confidence for male subjects is not significantly different between the two groups. This specifies that using sentiment scores did not have a significant effect on the confidence (in purchase decision) of male participants. The same analysis was conducted for female participants.

Table 17- ANCOVA to test H_3 on Females only

Tests of Between-Subjects Effects

Dependent Variable: Confidence

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	.741	1	.741	.629	.447
	Error	11.494	9.755	1.178[a]		
Group	Hypothesis	5.413	1	5.413	7.561	.010
	Error	21.477	30	.716[b]		
Education	Hypothesis	7.744	4	1.936	2.704	.049
	Error	21.477	30	.716[b]		
new_familiarity	Hypothesis	1.968	2	.984	1.374	.268
	Error	21.477	30	.716[b]		
Timefirst	Hypothesis	.077	1	.077	.107	.746
	Error	21.477	30	.716[b]		

The result shown in table 17 provides evidence that the confidence is significantly different between the two treatment groups considering females only. Therefore, the hypothesis (H_3) is supported for females. This confirms that female participants from the group using sentiment

scores reported different confidence levels than those of the group that did not use sentiment scores.

To find which group outperforms the other, in terms of confidence, we will use descriptive statistics table for this test provided in table 18. It is shown that the mean for confidence variable for the group using sentiment scores (group 0) is higher than that of the group that did not use sentiment scores. This result provides evidence to partially support H_3 that sentiment scores will help users to make decisions with a higher level of confidence. To be more exact, we may conclude **"sentiment scores help female users make decisions with a higher level of confidence**".

There are also some other interactions of education and gender with groups or familiarity and gender with groups that yield significance on confidence. However, we do not further analyze those, as the sample size would get too small due to multiple filtering and the results would not be reliable enough to report. But, the interactive effect of those control factors is worth further studies that will be in our future work plans.

Table 18- Descriptive Statistics for H_3

Descriptive Statistics

Dependent Variable: Confidence

Group	Education Level	Familiarity-group	Mean	Std. Deviation	N
Using SA	2	1.00	1.3848549	.	1
		3.00	.5825529	.30454818	3
		Total	.7831284	.47196947	4
	3	3.00	.8642125	.73629946	2
		Total	.8642125	.73629946	2
	5	1.00	.4135725	.	1
		2.00	-.5948555	.50514753	2
		3.00	.6704684	.	1
		Total	-.0264175	.72586928	4
	6	1.00	.6704684	.	1
		3.00	1.1911107	.	1
		Total	.9307895	.36814973	2
	Highest level of education	1.00	.3015529	.76568257	4
		2.00	.1532513	.73145537	2
		3.00	-1.0852028	.36330564	2
		Total	-.0822114	.85672102	8
	Total	1.00	.5250153	.67471615	7
		2.00	-.2208021	.67078680	4
		3.00	.3519175	.90003104	9
		Total	.2979578	.79555677	20
Not Using SA	2	2.00	.8036224	.82198678	2
		3.00	-1.2089446	.63106343	3
		Total	-.4039178	1.25823843	5
	3	1.00	-.4945581	.	1
		2.00	.4067219	.	1
		3.00	.6704684	.	1
		Total	.1942107	.61089479	3
	5	1.00	-.3466283	.36944853	3
		Total	-.3466283	.36944853	3
	6	2.00	.6704684	.00000000	2
		3.00	.1498260	.	1
		Total	.4969209	.30059300	3

	Highest level of education	1.00	-.8517514	1.51084676	4
		3.00	-.4314064	.	1
		Total	-.7676824	1.32186665	5
	Total	1.00	-.6176811	1.04030778	8
		2.00	.6709807	.44178138	5
		3.00	-.5396576	.90463565	6
		Total	-.2539206	1.00757831	19
Total	2	1.00	1.3848549	.	1
		2.00	.8036224	.82198678	2
		3.00	-.3131959	1.07667788	6
		Total	.1236583	1.12540032	9
	3	1.00	-.4945581	.	1
		2.00	.4067219	.	1
		3.00	.7996311	.53252297	3
		Total	.4622114	.67587068	5
	5	1.00	-.1565781	.48525368	4
		2.00	-.5948555	.50514753	2
		3.00	.6704684	.	1
		Total	-.1636507	.58158079	7
	6	1.00	.6704684	.	1
		2.00	.6704684	.00000000	2
		3.00	.6704684	.73629946	2
		Total	.6704684	.36814973	5
	Highest level of education	1.00	-.2750992	1.26868913	8
		2.00	.1532513	.73145537	2
		3.00	-.8672707	.45659472	3
		Total	-.3458541	1.06351921	13
	Total	1.00	-.0844228	1.04135736	15
		2.00	.2746328	.69801766	9
		3.00	-.0047126	.97957489	15
		Total	.0290940	.93565002	39

5. Discussion and Conclusions

Sentiment Analysis has become an active topic of research in data mining and text analytics literature. Various applications for sentiment analysis are proposed in different studies, yet not so many applications were empirically examined. It is really important for system designers to have an insight of who might benefit from their systems and how they may use their systems' outcomes to accomplish their tasks. Hence, an empirical investigation on the outcome of new systems yields to a better and most beneficial system design. The main purpose of this study has been to investigate the impact of sentiment analysis on purchase decision outcomes. The results provide empirical evidence on who benefits more from this technology as well as how this technology might affect purchase decision outcomes.

Since the investigation is on purchase decisions, marketers will also benefit from an evaluation of sentiment analysis technology on purchase decision outcomes before they commit to investments in this technology that, at a later time, might be found useless for their targeted market. In this study, we specifically focused on product reviews and how comparison aids such as sentiment scores may be used by consumers for information acquisition purposes hence affecting their final decision outcomes.

The investigation was performed in two different phases. In the first phase, we evaluated the potential effect of sentiment scores by comparing them to currently available numerical ratings, denoted as star ratings. Several studies confirmed the effect of star ratings on information acquisition processes and purchase decision outcomes. The results show a significant correlation between these two measures for various data sets from different domains.

The results from this part of our study suggest that sentiment scores, while being correlated with star ratings, are not able to detect the extreme sentiments. On the other hand, previous studies

indicate that extreme comments/reviews are more important and more useful for consumers while making a purchase decision [36]. Consequently, we decided to use an extended sentiment score for the second phase of our study. For that, a partially-linear mathematical equation was applied on the reported scores for each paragraph to extend the sentiment scores in a way that it could cover the extreme sentiment scores.

An experiment was designed and participants made purchase decisions for a set of specific products (similar in nature, but different in terms of reviews, brands and some general features) with and without sentiment scores on each paragraph. Huang et al.'s claim that sentiment analysis results- feature sentiments- can be used to make wiser decisions and to make those decisions significantly faster [39]. In our study, we evaluated the impact of "sentiment scores" rather than "sentiment feature" on decision outcomes. Decisions were evaluated from both effectiveness and efficiency. We hypothesized that sentiment scores will help users to make decisions significantly faster and with higher level of confidence. This is due to the potential impact of sentiment scores in information acquisition processes, finding extracted by first phase of our study.

Whilst this study did not confirm the effect of sentiment scores on time to make decision, it partially supported sentiment scores' impact on its relative subjective measure. Specifically, the results provide evidence that female consumers may use sentiment scores to make their purchase decisions with higher level of confidence.

The findings from this study can be used by users to improve their decision outcomes. Particularly, female consumers may use sentiment scores on long reviews that are not labeled with any other numerical ratings to improve their confidence in their decision. Marketers also, may benefit from these results. They may use sentiment scores while they are targeting female consumers to help them make a more confident decision on their purchase. They may also use

these numerical scores to accomplish certain marketing strategies, such as improving sales on a specific product targeting female consumers such as cosmetics or jewelries.

System designers may also benefit from the findings of our study. One potential suggestion would be to improve the design of sentiment analysis systems to provide more accurate results that are able to detect the extreme sentiments. This is due to the results from previous studies on star ratings as well as the results from this current study that investigated the impact of the extended sentiment scores on decision outcomes. Another suggestion would be to design systems that can be used as a built in sentiment analysis tool on review websites. Users may use those systems at their own discretions.

6. Limitations

The findings of this study have a number of important limitations. Some of these limitations are due to the design and data gathering while others can be referred to as common limitations of statistical analysis. One potential limitation can be due to the errors that might occur because of the use of statistical analysis with various assumptions. To tackle this problem we evaluated every single assumption and used only those variables that meet the assumption requirements. However, we still count for potential errors of evaluations and misleading findings.

One major limitation of this study is using an off the shelf sentiment analysis system that was not designed by a party who was involved in our study. This limitation has two main effects; one is the general flaw due to the original design of the system and the other is the notion of domain specificity that is suggested by almost all sentiment analysis studies. The system that was used in this study was not precisely designed for a specific domain; however this does not justify the domain specificity notion to be ignored.

Another possible limitation might be due to the fact that the decision in the experiment was made on a conceptual purchase decision and participants might not be as committed to make their decision as if they are doing it in a real experience. In addition, the limited amount of time provided to make their decision might be another limitation that caused the lack of support for our second hypothesis. This problem could be tackled if we were able to test our hypothesis in a real world case study or if we have designed the experiment in a way to mitigate that limitation e.g. providing such incentive to encourage a good decision.

Another limitation that we might refer to is that compared the impact of reviews on our participants, nevertheless only 70% of our participants reported that they used those reviews to

make their decision. Although the percentage of users is large enough, this observation might have affected the findings.

While we are aware of the potential limitations of our study, we are still confident that our results can be used by parties who benefit from it and further studies can be designed based on our reported findings.

7. Directions for future research

We have various roadmaps for future research. First is to further test our hypotheses with larger sample size to find other probable significant interactions using our control variables. As stated earlier, our current sample size was calculated with the Cohens' effect size of 0.65. If we consider a smaller effect size of 0.5 (which is considered as the exact medium effect assumption by Cohen), the power test using the same G*power software reports the total sample size of 210. With doubling our sample size we might be able to detect some more findings.

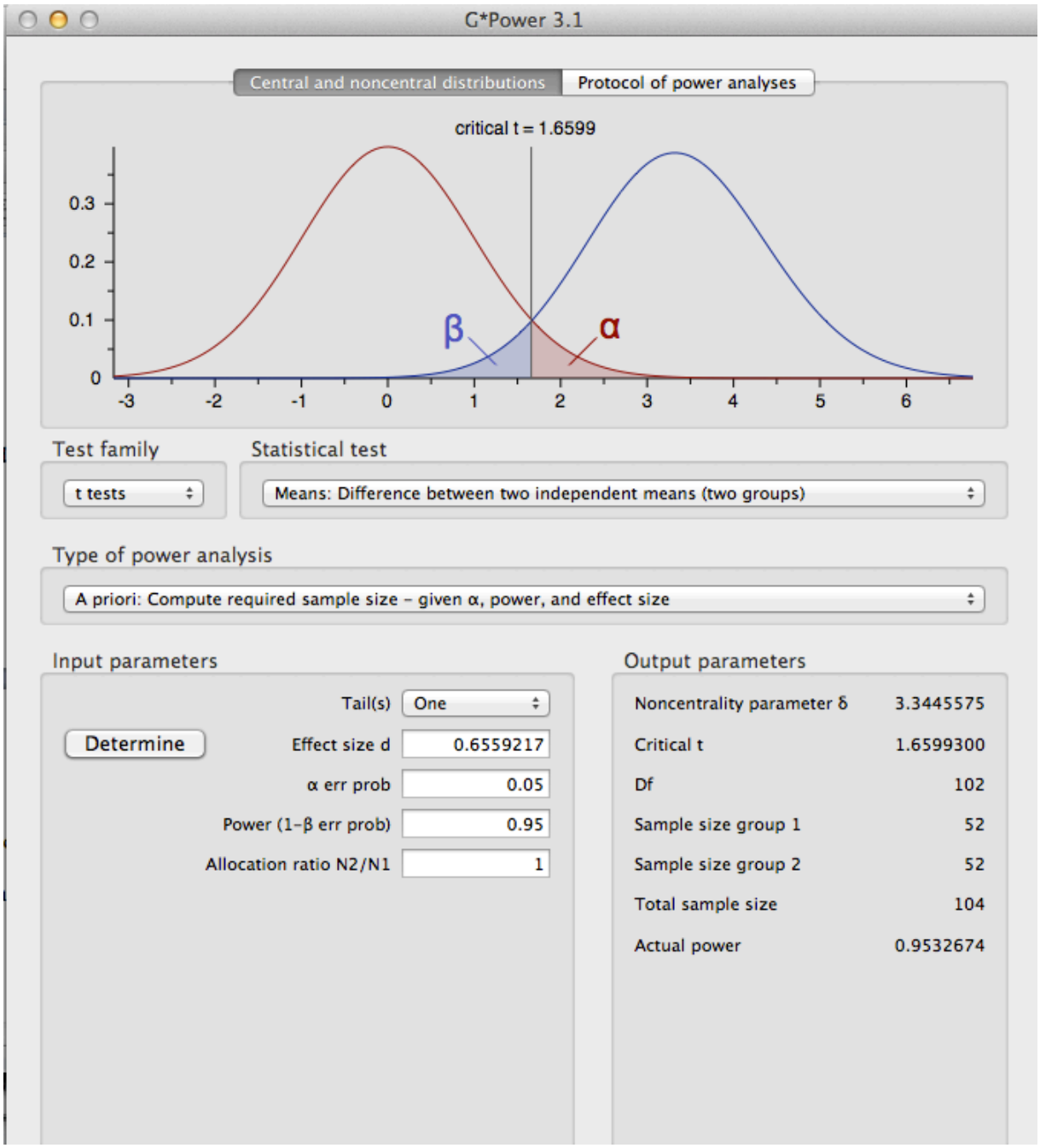
In this study, an extended version of sentiment scores, using a partially linear equation, was used. Sometimes the extent of negativity or positivity did not match the original content that while reading the text, the readers were disappointed and did not trust the scores. In our future work, we will use other mathematical techniques to create this sentiment extension in a way that provides more accurate results. This will be tested in our future pilot study.

Further we will investigate the impact of sentiment scores on other domains using the concepts of Task Technology Fit. In that we will examine the impact of sentiment scores on decision-making in other domains (e.g. doctor selection, travel plans, etc.).

Also, in our future research, we will consider the representation and visualization of sentiment scores and will test whether other representations of these scores might result in different effects on decision outcomes. For instance we might test whether the representation of the scores in the form of star ratings might be considered more cognitive by users and consequently might have stronger effect on decision outcomes.

8. Appendices

Appendix 1- Power test



Appendix 2- Chi-square test results

Doctor 1 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.831[a]	9	.128
Likelihood Ratio	14.169	9	.116
Linear-by-Linear Association	6.706	1	.010
N of Valid Cases	40		

a. 14 cells (87.5%) have expected count less than 5. The minimum expected count is .08.

Doctor 2 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	27.731[a]	12	.006
Likelihood Ratio	34.088	12	.001
Linear-by-Linear Association	23.466	1	.000
N of Valid Cases	62		

a. 15 cells (75.0%) have expected count less than 5. The minimum expected count is .03.

Doctor 3 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.733[a]	12	.318
Likelihood Ratio	15.224	12	.229
Linear-by-Linear Association	8.797	1	.003
N of Valid Cases	46		

a. 17 cells (85.0%) have expected count less than 5. The minimum expected count is .09.

Hotel 1 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	80.476[a]	12	.000
Likelihood Ratio	65.150	12	.000

Linear-by-Linear Association	42.103	1	.000
N of Valid Cases	119		

a. 13 cells (65.0%) have expected count less than 5. The minimum expected count is .13.

Hotel 2 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	46.713[a]	16	.000
Likelihood Ratio	47.684	16	.000
Linear-by-Linear Association	28.230	1	.000
N of Valid Cases	70		

a. 20 cells (80.0%) have expected count less than 5. The minimum expected count is .03.

Hotel 3 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18.057[a]	6	.006
Likelihood Ratio	18.545	6	.005
Linear-by-Linear Association	14.424	1	.000
N of Valid Cases	65		

a. 9 cells (75.0%) have expected count less than 5. The minimum expected count is .09.

Product 1 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	46.071[a]	20	.001
Likelihood Ratio	43.425	20	.002
Linear-by-Linear Association	14.241	1	.000
N of Valid Cases	53		

a. 27 cells (90.0%) have expected count less than 5. The minimum expected count is .09.

Product 2 Data Set

	Value	df	Asymp. Sig. (2-sided)

Pearson Chi-Square	27.649[a]	12	.006
Likelihood Ratio	32.564	12	.001
Linear-by-Linear Association	15.678	1	.000
N of Valid Cases	48		

a. 16 cells (80.0%) have expected count less than 5. The minimum expected count is .44.

Product 3 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	28.478[a]	12	.005
Likelihood Ratio	29.408	12	.003
Linear-by-Linear Association	15.424	1	.000
N of Valid Cases	46		

a. 17 cells (85.0%) have expected count less than 5. The minimum expected count is .13.

Product 4 Data Set

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.059[a]	12	.365
Likelihood Ratio	14.237	12	.286
Linear-by-Linear Association	1.672	1	.196
N of Valid Cases	46		

a. 18 cells (90.0%) have expected count less than 5. The minimum expected count is .13.

Appendix 3- Consent Form

Making Purchase Decision using online Reviews- Can Sentiment Analysis technology improve this process?

You are being invited to participate in a research study. Please read this Consent Form so that you understand what your participation will involve. Before you consent to participate, please ask any questions necessary to be sure you understand what your participation will involve.

Investigator

This research study is being conducted by Parisa Lak, MMSc student at Ted Rogers School of Management, Ryerson University, under the supervision of Dr. Ozgur Turetken, professor at TedRogers school of Management, ITM department, Ryerson University. If you have any questions or concerns about the research, please feel free to contact either Parisa Lak at parisa.lak@ryerson.ca or Dr.Turetken at turetken@ryerson.ca.

Purpose of study

We conducted this study to evaluate the usefulness of Sentiment Analysis technology- the automatic detection of the polarity of a document- in helping individuals, who use online reviews to make purchase decisions.

We intend to use the result of this study upon a thesis as well as submission to academic publication outlets, such as journals and conferences.

Description

If you volunteer to participate in this study, you will be asked to do the following things:

1- You are expected to make a purchase decision from three choices of professional cameras. The only source of information available for your decision-making is provided in a file containing pages of reviews. It is at your own discretion which part of the document is more important or useful to read.
2- You will need to answer questions regarding your age range and level of education as well your familiarity with professional cameras at the beginning of the questionnaire.
3- After you go over the information provided and, made your purchase decision, you will need to answer couple of questions about your decision. The questions are mainly regarding the information provided and how it helped you with your decision-making.

This whole process is expected not to take more than 1 hour. More instructions will be given during the session, but if you have any questions or concerns please ask any of the investigators.

Risks or Discomforts

The study is at minimal risk to the participants. In case of confusion or boredom participants may ask questions or raise concerns with the investigators at any time during the session.

Benefits of the Study

The result of this study will contribute to Information System and technology management literature by providing additional information regarding the value of investment on improvement of Sentiment Analysis technology and its usefulness for a specific task.

Confidentiality

All information gathered, including participant contact information and data gathered from the survey or through eventbrite website, will be kept confidential. The data will be securely stored on a password-protected server and the consent forms are going to be kept in a sealed envelope until the end of the study and then will be discarded. Only aggregate data will be used in any publications resulting from the study, and each participant will be recognized by his/her random control number, making it impossible to identify individual responses.

Costs and/or Compensation for Participation

This study won't take more than 60 minutes. In return, participants will receive $10 cash incentive and will be entered in a drawing for an Apple iPad with Retina display. The result of the draw will be announced via email to all participants. The winner will need to bring her/his control number to the investigators office located at TRS2-027 to claim her/his iPad.

Voluntary Nature of Participation

Participation in this study is voluntary. You can choose whether to be in this study or not. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind. If you choose to withdraw from this study you may also choose to withdraw your data from the study. You may also choose not to answer any question(s) and still remain in the study. Your choice of whether or not to participate will not influence your future relations with Ryerson University.

Questions about the Study

If you have any questions about the research now, please ask. If you have questions about the research later, you may contact Parisa Lak at parisa.lak@ryerson.ca or, Ozgur Turetken at turetken@ryerson.ca

If you have questions regarding your rights as a human subject and participant in this study, you may contact the Ryerson University Research Ethics Board for information.

Research Ethics Board

c/o Office of the Vice President, Research and Innovation

Ryerson University

350 Victoria Street

Toronto, ON M5B 2K3

416-979-5042

Agreement

Your signature below indicates that you have read the information in this agreement and have had a chance to ask any questions you have about the study. Your signature also indicates that you agree to be in the study and have been told that you can change your mind and withdraw your consent to participate at any time. You have been given a copy of this agreement.

You have been told that by signing this consent agreement you are not giving up any of your legal rights.

_____

Name of Participant (please print)

_____ _____

Signature of Participant Date

_____ _____

Signature of Investigator

Appendix 4 - Questionnaire

1. Background questions:

Gender

What is your Gender?

- Male
- Female

Education

What is the highest level of education you have completed?

- Some high school
- High school graduate
- Some college
- Trade/Technical/Vocational training
- College graduate
- Some post graduate work
- Postgraduate degree

Familiarity

Please specify to what extent you agree with below statements. 1= strongly disagree 7= strongly agree

I am familiar with DSLR cameras 1 2 3 4 5 6 7

2. Decision outcome questions:

Confidence

Please specify to what extent you agree with below statements. 1= strongly disagree 7= strongly agree

I am not confident that my choice of camera is the best out of the three

I have a strong feeling that the camera I chose is the best

I am confident about my choice of camera

3. Review usefulness questions (added for the second round data gathering):

Please specify to what extent you agree with below statements. 1= strongly disagree 7= strongly agree

Prior question

I will use camera reviews to make my decision

Post question

I intend to use reviews in my future decision-making task

I predict that I would use Reviews in my future decision-making

9. Bibliography

[1] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis Lectures on Human Language Technologies* 5, no. 1 (2012): 1-167.

[2] Buttle, Francis A. "Word of mouth: understanding and managing referral marketing." *Journal of strategic marketing* 6, no. 3 (1998): 241-254.

[3] Lu, Xianghua, Sulin Ba, Lihua Huang, and Yue Feng. "Promotional Marketing or Word-of-Mouth? Evidence from Online Restaurant Reviews." *Information Systems Research* (2013).

[4] Vessey, Iris, and Dennis Galletta. "Cognitive fit: An empirical study of information acquisition." *Information Systems Research* 2, no. 1 (1991): 63-84.

[5] Cox, Donald F., ed. *Risk taking and information handling in consumer behavior*. Boston: Division of Research, Graduate School of Business Administration, Harvard University, 1967.

[6] Hansen, Flemming. *Consumer choice behavior: A cognitive theory*. Vol. 10. New York: Free Press, 1972.

[7] Urbany, Joel E., Peter R. Dickson, and William L. Wilkie. "Buyer uncertainty and information search." *Journal of Consumer Research* (1989): 208-215.

[8] Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh, and Dwayne D. Gremler. "Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?." *Journal of interactive marketing* 18, no. 1 (2004): 38-52.

[9] Hatzivassiloglou, Vasileios, and Kathleen R. McKeown. "Predicting the semantic orientation of adjectives." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 174-181. Association for Computational Linguistics, 1997.

[10] Riloff E. and Wiebe J. 2003, "Learning extraction patterns for subjective expressions." *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 105–112, Sapporo, Japan.

[11] Pang B., Lee L., and Vaithyanathan S. 2002, "Thumbs up?: sentiment classification using machine learning techniques." *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. *Association for Computational Linguistics*.

[12] Paltoglou, G. and Thelwall, M. 2012. "Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media." *ACM Trans. Intell. Syst. Technol.* 3, 4, Article 66 (September 2012), 19 pages.

[13] Thelwall, Mike. "MySpace comments." *Online Information Review* 33, no. 1 (2009): 58-76.

[14] Thelwall, M., Buckley, K., Paltoglou, G., and Cai, D. 2010. "Sentiment strength detection in short informal text". *J. Amer. Soc. Inf. Sci. Technol. 61,* 12.

[15] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2, no. 1-2 (2008): 1-135.

[16] Kouloumpis, E. Wilson, Th. Moore J. 2011. "Twitter Sentiment Analysis: The Good the Bad and the OMG!" *Association for the Advancement of Artificial Intelligence*

[17] Jiang, L., Yu, M., Zhou, M., Liu, X. and Zhao, T. 2011. "Target-dependent twitter sentiment classification", *ACL,* 151–160.

[18] Liu, X., Li, K., Zhou, M., and Xiong, Z. 2011. "Collective semantic role labeling for tweets with clustering." In *IJCAI,* 1832–1837.

[19] Liu, X., Li, K., Zhou, M., and Xiong, Z. 2011. "Enhancing semantic role labeling for tweets using self-training." *AAAI*.

[20] Liu, X., Zhang, S., Wei, F., and Zhou, M. 2011. "Recognizing named entities in tweets." *ACL,* 359–367.

[21] Go, A., Bhayani, R., and Huang, L. 2009. "Twitter sentiment classification using distant supervision." *Technical report*

[22] Denis, P. and Sagot, B. 2009. "Coupling an annotated corpus and a morph syntactic lexicon for state-of-the-art POS tagging with less human effort." *Equip project ALPAGEINRIA and Université Paris 730*. Paris, France

[23] Hajic˘, J. 2000. "Morphological Tagging: Data vs. Dictionaries." *In Proceedings of ANLP' 00,* pages 94–101, Seattle, WA, USA.

[24] Pirolli, Peter, and Stuart Card. "Information foraging." *Psychological review* 106, no. 4 (1999): 643.

[25] Owsley S., Sood S., and Hammond, K. J. 2006, "Domain specific affective classification of documents." *In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 181–183, Stanford, CA, USA.

[26] Mudambi, Susan M., and David Schuff. "What makes a helpful online review? A study of customer reviews on Amazon. com." *MIS quarterly* 34, no. 1 (2010): 185-200.

[27] Stigler, George J. "The economics of information." *The journal of political economy* 69, no. 3 (1961): 213-225.

[28] P. Nelson "Information and Consumer Behavior," *Journal of Political Economy* (78:20), 1970, pp. 311-329.

[29] E. Johnson, and J. Payne "Effort and Accuracy in Choice," *Management Science* (31:4), 1985, pp. 395-415.

[30] D. Godes, and D. Mayzlin. "Using online conversations to study word-of-mouth communication." *Marketing Science* (23:4), 2004, pp. 545-560.

[31] N. Kumar, and I. Benbasat "The Influence of Recommendations on Consumer Reviews on Evaluations of Websites," *Information Systems Research* (17:4), 2006, pp. 425-439.

[32] Nelson, P. 1970. "Information and Consumer Behavior," *Journal of Political Economy* (78:20), pp. 311-329.

[33] Chevalier, Judith A., and Dina Mayzlin. *The effect of word of mouth on sales: Online book reviews*. No. w10148. National Bureau of Economic Research, 2003.

[34] P. Todd, and I. Benbasat "The Use of Information in Decision Making: An Experimental Investigation of the Impact of Computer-Based Decision Aids," *MIS Quarterly* (16:3), 1992, pp.373-393

[35] R. Poston, and C. Speier, "Effective Use of Knowledge Management Systems: A Process Model of Content Ratings and Credibility Indicators," *MIS Quarterly* (29:2), 2005, pp. 221-244

[36] P. Pavlou, and A. Dimoka "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research* (17:4), 2006, pp. 392-414.

[37] C. Forman, A. Ghose, B. Wiesenfeld. "Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research* (19:3), 2008, pp. 291-313.

[38] A. E. Crowley, and W. D. Hoyer "An Integrative Framework for Understanding Two-Sided Persuasion," *Journal of Consumer Research* (20:4), 1994, pp. 561-574.

[39] Huang, Shih-Wen, Pei-Fen Tu, Wai-Tat Fu, and Mohammad Amamzadeh. "Leveraging the crowd to improve feature-sentiment analysis of user reviews." In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 3-14. ACM, 2013.

[40] B. Liu, "Sentiment analysis and opinion mining." *Synthesis Lectures on Human Language Technologies* (5: 1), 2012, pp. 1-167.

[41] Cambria, Erik, Yangqiu Song, Haixun Wang, and Newton Howard. "Semantic multi-dimensional scaling for open-domain sentiment analysis." (2013): 1-1.

[42] Rosas, Veronica, Rada Mihalcea, and L. Morency. "Multimodal Sentiment Analysis of Spanish Online Videos." (2013): 1-1.

[43] G. Paltoglou, and M. Thelwall. "Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media." *ACM Transactions on Intelligent Systems and Technology* (TIST) (3:4), 2012, p.66.

[44] Liebmann, Michael, Michael Hagenau, and Dirk Neumann. "Information Processing in Electronic Markets: Measuring Subjective Interpretation Using Sentiment Analysis." (2012).

[45] Zhao, Jichang, Li Dong, Junjie Wu, and Ke Xu. "Moodlens: an emoticon-based sentiment analysis system for chinese tweets." In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1528-1531. ACM, 2012.

[46] O. Turetken, O., and L. Olfman, "Introduction to the Special Issue on Human-Computer Interaction in the Web 2.0 Era," *AIS Transactions on Human-Computer Interaction* (5:1), 2013 pp. 1-5.

[47] Goodhue, Dale L., and Ronald L. Thompson. "Task-technology fit and individual performance." *MIS quarterly* (1995): 213-236.

[48] Goodhue, D.L. (1995), "Understanding User Evaluations of Information Systems", *Management Science 41.12,* 1827-1844.

[49] Joachims, Thorsten. "Making large scale SVM learning practical." (1999).

[50] Russom, Philip. "Big data analytics." *TDWI Best Practices Report*, Fourth Quarter (2011).

[51] Meade, Phillip T., and Luis Rabelo. "The technology adoption life cycle attractor: Understanding the dynamics of high-tech markets." *Technological Forecasting and Social Change* 71, no. 7 (2004): 667-684.

[52] Goodhue, Dale. "The model underlying the measurement of the impacts of the IIC on the end-users." *Journal of the American Society for Information Science* 48, no. 5 (1997): 449-453.

[53] Tan, Joseph KH, and Izak Benbasat. "Processing of graphical information: A decomposition taxonomy to match data extraction tasks and graphical representations." *Information Systems Research* (1990): 416-439.

[54] Vessey, Iris. "Cognitive Fit: A Theory‐Based Analysis of the Graphs Versus Tables Literature*." *Decision Sciences* 22, no. 2 (1991): 219-240.

[55] Jarvenpaa, Sirkka L. "The effect of task demands and graphical format on information processing strategies." *Management Science* 35, no. 3 (1989): 285-303.

[56] Dishaw, Mark T., and Diane M. Strong. "Extending the technology acceptance model with task–technology fit constructs." *Information & Management* 36, no. 1 (1999): 9-21.

[57] Goodhue, Dale L. "Development and Measurement Validity of a Task‐Technology Fit Instrument for User Evaluations of Information System." *Decision Sciences* 29, no. 1 (1998): 105-138.

[58] Dishaw, Mark T., and Diane M. Strong. "Extending the technology acceptance model with task–technology fit constructs." *Information & Management* 36, no. 1 (1999): 9-21.

[59] Huang, Zan, Hsinchun Chen, Fei Guo, Jennifer J. Xu, Soushan Wu, and Wun-Hwa Chen. "Expertise visualization: An implementation and study based on cognitive fit theory." *Decision Support Systems* 42, no. 3 (2006): 1539-1557.

[60] Dickson, Gary W., Gerardine DeSanctis, and DOROTHY J. McBride. "Understanding the effectiveness of computer graphics for decision support: a cumulative experimental approach." *Communications of the ACM* 29, no. 1 (1986): 40-47.

[61] Staples, D. Sandy, and Peter Seddon. "Testing the technology-to-performance chain model." *Journal of Organizational and End User Computing (JOEUC)* 16, no. 4 (2004): 17-36.

[62] Granados, Nelson, Alok Gupta, and Robert J. Kauffman. "Research Commentary—Information Transparency in Business-to-Consumer Markets: Concepts, Framework, and Research Agenda." *Information Systems Research* 21, no. 2 (2010): 207-226.

[63] Comrey, Andrew Laurence, and Howard Bing Lee. *A first course in factor analysis*. Routledge, 1992.

[64] Smith, Eric A., and Bruce Winterhalder, eds. *Evolutionary ecology and human behavior*. Transaction Books, 1992.

[65] Stephens, David W. *Foraging theory*. Princeton University Press, 1986.

[66] S. S. Sundar, S. Knobloch-Westerwick and M. R. Hastall, "News Cues: Information Scent and Cognitive Heuristics," *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 3, 2007, pp. 366- 378.

[67] Khapre, Shailesh, and MS Saleem Basha. "Advancement in Information Foraging Theory." *Intelligent Information Management* 4, no. 6 (2012): 383-389.

[68] Pirolli, Peter. "Computational models of information scent-following in a very large browsable text collection." In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pp. 3-10. ACM, 1997.

[69] Pirolli, Peter, James Pitkow, and Ramana Rao. "Silk from a sow's ear: extracting usable structures from the Web." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 118-125. ACM, 1996.

[70] Brunswik, Egon. *Perception and the representative design of psychological experiments*. University of California Pr, 1956.

[71] Anderson, John R. "The adaptive nature of human categorization." *Psychological Review* 98, no. 3 (1991): 409.

[72] J Anderson, John R., ed. *The adaptive character of thought*. Psychology Press, 1990.

[73] McFadden, Daniel. *Modelling the choice of residential location*. Institute of Transportation Studies, University of California, 1978.

[74] Payne, John W. "Contingent Decision Behavior: A Review and Discussion of Issues." *psychological bulletin*, 92, 2 (1982), 382-402.

[75] Christensen-Szalanski, Jay JJ. "A further examination of the selection of problem-solving strategies: The effects of deadlines and analytic aptitudes." *Organizational Behavior and Human Performance* 25, no. 1 (1980): 107-122.

[76] Creyer,E.H., J.R.Bettman, and J.W. Payne, "The Impact of Accuracy and Effort Feedback and Goals on Adaptive Decision Behavior," *J. BehavioralDecisioniMakinig*,3, 1 (1990), 1-16.

[77] DiStefano, Christine, Min Zhu, and Diana Mindrila. "Understanding and using factor scores: Considerations for the applied researcher." *Practical Assessment, Research & Evaluation* 14, no. 20 (2009): 1-11.

[78] Gînscă, Alexandru-Lucian, Emanuela Boroş, Adrian Iftene, Diana TrandabĂţ, Mihai Toader, Marius Corîci, Cenel-Augusto Perez, and Dan Cristea. "Sentimatrix: multilingual sentiment analysis service." *In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 189-195. Association for Computational Linguistics, 2011.

[79] Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences." *Behavior research methods* 39, no. 2 (2007): 175-191.

[80] Cohen, Jacob. "Statistical power analysis." *Current directions in psychological science* 1, no. 3 (1992): 98-101.

[81] Schindler, Robert M., and Barbara Bickart. "Published word of mouth: Referable, consumer-generated information on the Internet." *Online consumer psychology: Understanding and influencing consumer behavior in the virtual world* (2005): 35-61.

[82] Schiffman, Leon G., and Leslie Lazar Kanuk. "Consumer Behavior, 7th." (2000).

[83] Goldsmith, Ronald E., and David Horowitz. "Measuring motivations for online opinion seeking." *Journal of interactive advertising* 6, no. 2 (2006): 1-16.

[84] Keppel, Geoffrey. *Design and analysis: A researcher's handbook* . Prentice-Hall, Inc, 1991.

[85] D'Alonzo, Karen T. "The Johnson-Neyman procedure as an alternative to ANCOVA." *Western journal of nursing research* 26, no. 7 (2004): 804-812.

[86] MacCallum, Robert C., and Michael W. Browne. "The use of causal indicators in covariance structure models: some practical issues." *Psychological bulletin* 114, no. 3 (1993): 533.

Davis, Fred D. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." *MIS quarterly* (1989): 319-340.

[87] Lee, Yang W., Diane M. Strong, Beverly K. Kahn, and Richard Y. Wang. "AIMQ: a methodology for information quality assessment." *Information & management* 40, no. 2 (2002): 133-146.

[88] Venkatesh, Viswanath, and Fred D. Davis. "A theoretical extension of the technology acceptance model: four longitudinal field studies." *Management science* 46, no. 2 (2000): 186-204.

[89] O'Connor, Annette M. "Validation of a decisional conflict scale." *Medical decision making* 15, no. 1 (1995): 25-30.

[90] Van der Heijden, Hans. "User acceptance of hedonic information systems." *MIS quarterly* (2004): 695-704.

[91] Y. Qiang, R. Law, and B. Gu. "The impact of online user reviews on hotel room sales." *International Journal of Hospitality Management* (28:1), 2009, pp. 180-182.

[92] Lexalytics, "Lexalytics Web Demo," www.lexalytics.comlwebdemo. [accessed: 4 June 2013]

[93] Dellarocas, C., Zhang, M., and Awad, N. F. 2007. "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures", *Journal of Interactive Marketing* Volume 21 (4), Fall 2007, 23-45

[94] Reinstein, D., and Snyder, C. 2005. „The influence of expert reviews on consumer demand for experience goods: A case study of movie critics", *Journal of Industrial Economics* 53(1): 27-51.

[95] Chevalier, J., Mayzlin, D. 2006. "The effect of word of mouth online: Online book reviews". *Journal of Marketing Research* 43(3), 345-354.

[96] Yatani, K., Novati, M., Trusty, A., and Truong, K. N. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. *In Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems, CHI '11, ACM (New York, NY, USA, 2011), 1541–1550.

[97] M. Hu and B. Liu, "Mining and summarizing customer reviews," *in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, Washington, 2004.

[98] P. Carvalho, L. Sarmento, J. Teixeira, and M. Silva, "Liars and saviors in a sentiment annotated corpus of comments to political debates," *in Proceedings of the Association for Computational Linguistics (ACL 2011)*, Portland, OR, 2011.

[99] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," *in Conference on Empirical Methods in Natural Language Processing (EMNLP-03),* Sapporo, Japan, 2003, pp. 129–136.

[100] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large- scale news analysis," *in String Processing and Information Retrieval (SPIRE 2005),* 2005.

[101] Johan Bollen, Alberto Pepe, and Huina Mao. "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena". *5th ICWSM,* 2011.

[102] Y. Niu, X. Zhu, J. Li, and G. Hirst, "Analysis of polarity information in medical text," *in Proceedings of the American Medical Informatics Association 2005 Annual Symposium,* 2005.

[103] V. Hatzivassiloglou and J. Wiebe, "Effects of adjective orientation and grad- ability on sentence subjectivity," *in Proceedings of the International Conference on Computational Linguistics (COLING*), 2000.

[104] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *in Proceedings of the Association for Computational Linguistics (ACL),* pp. 417–424, 2002.

[105] Yang, Kiduk, Ning Yu, Alejandro Valerio, and Hui Zhang. "*WIDIT* in TREC 2006 Blog Track." In TREC. 2006.

[106] Engelberg, Joseph. "Costly information processing: Evidence from earnings announcements." In *AFA 2009 San Francisco Meetings Paper*. 2008.

[107] Granados, Nelson, Alok Gupta, and Robert J. Kauffman. "Research Commentary—Information Transparency in Business-to-Consumer Markets: Concepts, Framework, and Research Agenda." *Information Systems Research* 21, no. 2 (2010): 207-226.

[108] Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10 (2010): 178-185.

[109] Thomas, Matt, Bo Pang, and Lillian Lee. "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts." *In Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 327-335. Association for Computational Linguistics, 2006.

[110] Vessey, Iris. "The effect of information presentation on decision making: A cost-benefit analysis." *Information & Management* 27, no. 2 (1994): 103-119.

[111] Drèze, Jean, and Nicholas Stern. "The theory of cost-benefit analysis." *Handbook of public economics* 2 *(*1987): 909-989.

[112] Pirolli, Peter LT. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, 2007.