Star Ratings versus Sentiment Analysis - A Comparison of Explicit and Implicit Measures of Opinions

Parisa Lak
Ryerson University
parisa.lak@ryerson.ca

Ozgur Turetken
Ryerson University Management
turetken@ryerson.ca

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. 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. In this study, we compare sentiment analysis results with star ratings in three different domains to explore the promise of this technique.

1. Introduction

Opinions are central to many human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are, to a degree, dependent upon how others see and evaluate the world [1]. Therefore by analyzing opinions, not only can one predict certain behaviors of individuals expressing those opinions, but also those of others exposed to those opinions.

The rapid increase in the volume of Internet users and the growth of web 2.0 popularity among those users gave rise to massive collections of user-generated content [2]. The easy and ubiquitous access to these collections facilitates communication between individuals often from different cultures and different backgrounds, regardless of their geographical and demographical differences. This encourages some to freely state their opinions on a variety of issues, and others to use those opinions in their decisions. This valuable data can also be gathered and used by managers to evaluate the products and services offered by their organizations from the consumer's point of view.

A popular platform that facilitates such exchanges is a review website. Many consumers use reviews posted by other consumers before making their purchase decisions. A typical review site allows its users to indicate their opinions about a product or service in an open-ended form while also providing an opportunity to summarize these opinions through star ratings. Star ratings have been shown to serve as valid cues of content of a long review [3]. Research has also established the helpfulness of star ratings for consumers to select the most useful comments to read in more detail [4, 5, 6, 7].

Meanwhile, user generation of content in other forms such as blog posts, tweets, and news is ongoing. Most of this content is unstructured and lacks the information cues provided by star ratings. Nevertheless this information can be invaluable if managed effectively. The problem of managing unstructured or semi-structured text dates back to the pre-internet era (e.g., [8]), and data scientists and developers have long been working on various text analytic techniques to tackle this problem. Yet the ubiquitous nature of the Internet and the way it facilitates user generated content has made it essential to use text analytical tools and techniques to leverage these information sources. One such technique that has gained recent popularity is sentiment analysis.

Sentiment Analysis uses various classification techniques to identify the tone of a given piece of text. It indicates whether the 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. For example, sentiment analysis of a large amount of user feedback on a specific product or service helps managers to obtain a quick understanding of the response to their offerings. Likewise, it can help politicians to determine whether their campaign messages are resonating with likely voters. In this

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.106

IEEE computer society

study, we study the effectiveness of sentiment analysis on the comments from various product and service review websites. We compare the results of sentiment analysis with star ratings. Our objective is to ascertain whether sentiment analysis results can be used as an alternative to star ratings when such ratings are available, and more importantly, as a substitute when ratings are unavailable as in the case of blog posts and news. Therefore the research question that we address in this study is:

How comparable to star ratings are sentiment analysis results of reviews/comments?

Star ratings and sentiment analysis scores should be fairly close if comments about a product are consistent with 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. This will help use the findings of research on the usefulness of star ratings as a justification for the investment in the development of sentiment analysis techniques and tools, because sentiment analysis of reviews without ratings can then be used as a cue on which of those reviews are more useful to read. Further, sentiment analysis results can be used to decide not only which reviews to read, but also which parts of a long review to pay more attention to.

The rest of this paper is organized as follows. In the next section we review prior work on sentiment analysis as well as studies on star rating and its impact on choosing useful reviews for a purchase decision. We then explain our methodology, describing the data used in our comparisons as well as the sentiment analysis tool used for the analysis of that data. Then we report on the results of the statistical analysis comparing star ratings and sentiment analysis scores. We discuss these results, draw conclusions and identify directions for future research.

2. Background

Consumers usually make purchase decisions with a level of uncertainty. They use review websites as good sources of information to reduce this uncertainty [4]. However, information gathering comes with a cost, which is the time (and sometimes financial resources) spent to gather, analyze, and comprehend that information. Individuals normally are aware of the trade-offs between the perceived costs and benefits of search [9]. 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 [10]. For a wide range of choices, consumers recognize that there are tradeoffs between effort and accuracy [11].

"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)"[12]. With the extensive use of interactive web and the 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 [13] 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 rank 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 [14] in that it is important to discern which reviews are the most useful and actually able to reduce consumers' purchase uncertainty. According to Chevalier and Mayzlin [15], star ratings provide an excellent opportunity to measure the valence of comments without analyzing the comments themselves.

Consumers can use decision and comparison aids [16] and numerical content ratings (such as star ratings) [3] to conserve cognitive resources and reduce energy expenditure to acquire information, but also to ease or improve the purchase decision process [4]. The star rating has been shown to serve as a cue for the review content [3].

There have been numerous studies on consumer's perception of usefulness of positive and negative reviews. For instance, in [5] 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. [6] found that for books, moderate reviews (3 stars) were less helpful than extreme reviews. However, Crowley and Hoyer [7] 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.

However, 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 should one read given the limited resources (time) 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 the variety of data. One text analytics technology with promise to address this problem is sentiment analysis. Sentiment analysis tools use different classification techniques to determine the polarity of a piece of text and summarize this polarity through a quantitative measure. Various applications have been developed based on this basic principle. Some of those applications are listed in table 1. For instance, Huang et al. [17] indicate that sentiment results can be used to make wiser decisions and to make those decisions significantly faster. Liu [1] specifies various applications of sentiment analysis from the evaluation of consumer products, services, healthcare, and financial services to the analysis of social events and political elections. 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.

Table 1. Summary of past research using sentiment analysis

Author	Year	Use of Sentiment Analysis technology
García et al. [18]	2012	▪ Analyzing Spanish reviews
Cambria et al. [19]	2013	▪ Distilling useful information from unstructured data
Rosas et al. [20]	2013	▪ Branding and product analysis ▪ Tracking sentiment timelines in on-line forums and news ▪ Analysis of political debates ▪ Question answering ▪ Conversation summarization
Huang et al. [17]	2013	▪ Make wiser decisions ▪ Make decisions significantly faster
Paltoglu et al. [21]	2012	▪ 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. [22]	2013	▪ Resource allocation of E-commerce ▪ Financial prediction (difference between analyst and investors decisions)
Jichang et al. [23]	2012	▪ Understanding user behaviors

3. Methodology

3.1. Description of Data

To conduct this 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.

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 [24]. Therefore, the second domain chosen for analysis was hotel reviews.

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.

The data were gathered using tripadvisor (http://www.tripadvisor.com/) for hotel reviews and

RateMDs (http://www.ratemds.com/) for doctor reviews. 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. 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. 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. For doctors' reviews, three family doctors, from Toronto, were randomly selected from the website. 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 rating 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.

3.2. Sentiment Analysis Tool

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[1]) 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.[2] 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[3] and sentiment 140[4]. Sentiment 140 is basically developed for tweets and is not able to analyze 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 are disqualified for our study. Sysomos [5], viralheat [6], lithium[7], Gravity[8] and Datasift[9] are some examples of such software.

To support our claim that Lexalytics is an adequate choice for our experiment, we compared the results from Lexalytics to those from a system similar to Lexalytics in the way it can handle long texts and create a single sentiment score. According to [10] [25], Lymbix (along with Lexalytics) is a state of the art sentiment analysis tool. Lybix can analyze documents that are longer than tweets, but still limits the number of words to 20,000. This drawback made this system not to be our first choice tool in our study. However, this system is able to provide numerical scores (from -10 to 10) rather than binary (positivity and negativity) reports. Hence, we were able to compare the results from the two systems. This comparison was made to test the accuracy of Lexalytics compared to another state of the art 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 table 2a, Lymbix was not able to analyze 3 comments out of our sample of 88, because they contained more than 20,000 words. The correlation analysis conducted on the remaining 85 comments however (see table 2b) shows that the results from the two systems are highly correlated, which further confirms that Lexalytics is a good representative for systems that perform sentiment analysis.

Lexalytic 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

[1] http://www.lexalytics.com/demo
[2] http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_1/
[3] http://sentistrength.wlv.ac.uk/
[4] http://www.sentiment140.com/

[5] http://www.sysomos.com/
[6] https://www.viralheat.com/
[7] http://www.lithium.com/
[8] http://www.gravity.com/
[9] http://datasift.com/
[10] http://www.lymbix.com/

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, awful) [26]. 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.

Table 2. Lexalytics and Lymbix comparison

a. Descriptive Statistics

	Mean	Std. Deviation	N
Lexalytics	.346080	.2868738	88
Lymbix	2.916482	4.2470722	85

b. Correlations

		Lexalytics	Lymbix
Lexalytics	Pearson Correlation	1	.328**
	Sig. (2-tailed)		.002
	N	88	85
Lymbix	Pearson Correlation	.328**	1
	Sig. (2-tailed)	.002	
	N	85	85

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

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 [26].

3.3. Analysis of the Content

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 a 1 to 5 score and rounded them to their nearest integer values to make them compatible with star ratings.

To compare each sentiment analysis score with its corresponding star rating 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 from 1 to 5) for a comment is the same as the corresponding star rating.

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).

Due to space limitations, we include only two representative cross tabulations (one with significantly different distributions, and one without significantly different distributions) of sentiment analysis scores and star ratings (see Appendix A). 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 B. 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 3 displays the results. As seen in the table, 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 3- 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)

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 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 paper suggest 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 paper.

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 an important domain 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.

5. References

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

[2] 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.

[3] 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

[4] S. M. Mudambi, and D. Schuff. "What makes a helpful online review? A study of customer reviews on Amazon. com." MIS Quarterly (34:1), 2010, pp. 185-200.

[5] 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.

[6] 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.

[7] 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.

[8] G. Salton "Mathematics and Information Retrieval", Journal of Documentation (35:1), 1979, pp.1-29

[9] G. J. Stigler "The Economics of Information," Journal of Political Economy (69:3), 1961, pp. 213-225.

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

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

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

[13] 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.

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

[15] 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.

[16] 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

[17] Huang, Shih-Wen, Pei-Fen Tu, Wai-Tat Fu, and Mohammad Amanzadeh. "Leveraging the Crowd to Improve Feature-Sentiment Analysis of User Reviews." (2013).

[18] A. García, S. Gaines, and M. T. Linaza. "A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain." In ENTER 2012 Idea Exchange. 19th International Conference on Information Technology and Travel & Tourism ENTER 2012'eTourism Present and Future Services and Applications', Helsingborg, Sweden, 24-27 January 2012., (10:2) Texas A & M University Press, 2012.

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

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

[21] 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.

[22] M. Liebmann, M. Hagenau, and D.Neumann "Information Processing in Electronic Markets: Measuring Subjective Interpretation Using Sentiment Analysis." (2013).

[23] Z. Jichang, L. Dong, J. Wu, and K. 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, ACM, 2012, pp. 1528-1531

[24] 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.

[25] 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

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

Appendix-A
Cross Tabulations of Normalized Sentiment Analysis Scores (SA) and Star Ratings (SR)

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

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

Appendix-B
Chi-Square Tests

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.