Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (eWOM) communication

Chuanyi Tang · Lin Guo

Abstract Text-based electronic word-of-mouth (eWOM) communication has increasingly become an important channel for consumers to exchange information about products and services. How to effectively utilize the enormous amount of text information poses a great challenge to marketing researchers and practitioners. This study takes an initial step to investigate the validities and usefulness of text mining, a promising approach in generating valuable information from eWOM communication. Bilateral data were collected from both eWOM senders and readers via two web-based surveys. Results provide initial evidence for the validity and utility of text mining and demonstrate that the linguistic indicators generated by text analysis are predictive of eWOM communicators' attitudes toward a product or service. Text analysis indicators (e.g., Negations and Money) can explain additional variance in eWOM communicators' attitudes above and beyond the star ratings and may become a promising supplement to the widely used star ratings as indicators of eWOM valence.

Keywords Electronic word-of-mouth · Text mining · Linguistic indicator · Attitude

1 Introduction

Electronic word-of-mouth (eWOM) is defined as information communication among consumers about a product or company via the Internet (Henning-Thurau et al. 2004). Unlike the traditional WOM dominated by face-to-face or telephone communication,

C. Tang (✉)
Department of Marketing, College of Business & Public Administration, Old Dominion University, Norfolk, VA 23529-0220, USA
e-mail: ctang@odu.edu

L. Guo
Department of Marketing, Peter T. Paul College of Business and Economics, University of New Hampshire, Durham, NH 03824, USA
e-mail: Lin.Guo@unh.edu

eWOM is characterized as text-based with a sender typing a message on the Internet and a number of readers reading the text message on the Internet. In recent years, eWOM has become an increasingly important channel by which consumers exchange information about products and services. Jansen (2010) shows that over 50 % of adult consumers used eWOM to research product information and approximately 24 % of consumers posted online reviews. The enormous amount of eWOM messages posted on the Internet poses opportunities as well as challenges to marketing researchers and practitioners. On the one hand, initiated by consumers in natural situations, eWOM provides a goldmine of voluminous, authentic consumer information. On the other hand, digging into the goldmine and actually utilizing the enormous quantity of unstructured consumer information can be challenging (Lehto et al. 2007).

Capable of processing a large amount of text data effectively, text mining techniques have shown great potential in overcoming the challenge and driving valuable insights from the text-based eWOM communication. However, as a relatively new method, little is known regarding whether the text mining method is valid and how much additional value it can contribute to eWOM research. In an attempt to fill this gap, this study takes an initial step to examine the validity and utility of text mining in studying eWOM communication. Specifically, this study focuses on consumers' attitudes and aims to (1) identify the linguistic indicators generated by text mining that are significantly correlated with eWOM communicators' attitudes toward a product/service, (2) evaluate the predictive validity of text mining indicators on eWOM communicators' attitudes, (3) examine the correlations between text mining indicators of emotion word uses and consumers' corresponding self-reported measures, and (4) compare the predictive power of text mining indicators on eWOM communicators' attitudes with those of the star ratings.

2 Background on text mining

Text mining, also called text analysis, text analytics, text data mining, automatic text analysis, and computer-based text analysis, is the analysis of text data in order to discover hidden patterns, traits, and relationships. Beyond the content and linguistic analysis of texts, text mining often includes extended functions, such as database building, statistical analysis, and outcome visualization. Despite the different terminologies and technologies used, the common aim of text mining is to transfer the unstructured text information into structured information, which can then be analyzed with traditional data mining and statistical techniques.

Traditionally, text analysis has been used in other disciplines such as psychology to analyze the content of written information and to predict psychological states and behaviors of individuals (e.g., Bantum and Owen 2009). Some initial evidence regarding the validity and utility of text mining methods has been provided in psychology (e.g., Alpers et al. 2005; Kahn et al. 2007; Pennebaker et al. 1997). However, because most of the psychological studies employing text mining were conducted in the context of therapeutic discourse and health-related problems (e.g., emotional writing of trauma and cancer narratives), which involved intense emotions, it is unclear whether the text mining method could be safely employed to capture eWOM information in consumers' normal lives.

Text mining can be conducted from a qualitative approach, a quantitative approach, or a combination of both. Qualitative analysis generates non-numerical information and is commonly used to understand the underlying reasons and motivations, to identify the underling themes or relationships, and to develop hypotheses (for a good example, see Kozinets 2002). In contrast, the quantitative approach treats texts as objective data and generates numerical data. This approach is more widely used (e.g., Liu 2006; Sirdhar and Srinivasan 2012) largely because its output can be directly used in conventional statistical analysis.

Typical text mining tasks include information extraction, text categorization, text clustering, document summarization, and association analysis (e.g., Pennebaker et al. 2003; Gupta and Lehal 2009; Ramanathan and Meyyappan 2013). Information extraction is the most basic function of text mining. This technique can be used to identify the important market terms and themes such as brand names and product features from online forums and social media. Categorization is used to classify and assign texts into predefined categories or themes. In categorization, computer programs often treat a text as a bag of words and count the word frequencies. For example, the words "trouble" and "anger" might be assigned to the category of "negative emotions." This approach is widely used to identify communicators' attitudes or emotional states in sentiment analysis. Clustering is a technique used to group similar documents. It differs from categorization in that it does not use pre-defined categories, but instead clusters documents in real time. Association analysis is to find associations for a given term based on counting co-occurrence frequencies. For example, if one brand name always appears in social media with positive adjectives, it indicates that consumers may have a positive image of this brand. Summarization is to summarize the important concepts in texts while reducing the length and detail of a document. Based on a large collection of text documents, summarization can also help to identify the market trends, such as the changes of consumer preferences over time. In addition, text mining programs often have other additional functions such as visualization, in which a graphical information representation can be created.

With the advances in computer technology, many computer-based text mining tools have been developed, which makes text mining easier. In Table 1, we provide an overview of some major text mining tools. Recently, a couple of innovative studies have employed text mining tools to obtain valuable information from eWOM. For example, Aggarwal et al. (2009) used lexical semantic analysis, a text mining approach, to assess brand positions relative to competitors. They first collected web content related to their targeted brands through Google. They then conducted lexical semantic analysis to examine the co-occurrence of brand names and key adjective descriptors. For example, if Sony often appears on webpages with the adjective "reliable", it indicates that "reliable" is a main characteristic of Sony.

3 Research questions

As shown in Fig. 1, we propose that eWOM communication begins when an eWOM sender develops attitudes toward a product/service based on his/her consumption experience(s). His/her attitudes are then incorporated into a text review of the

Table 1 Text mining tools

Text miner	Vendor	Description	URL
SAS Text Analytics	SAS Institute Inc.	Functions include information extraction, categorization, clustering, sentiment analysis, association analysis, and summarization	http://www.sas.com/text-analytics/
SPSS Modeler Premium	IBM SPSS	Integrates unstructured text data into modeling. Functions include information extraction, clustering, association analysis, categorization, summarization, and visualization	http://www.spss.com.hk/software/modeler/index.htm?tab=1
Nvivo	QSR International	Supports qualitative and mixed-method research. Functions include information extraction, clustering, categorization, coding, and information visualization etc.	http://www.qsrinternational.com/products_nvivo.aspx
Linguistic Inquiry and Word Count	Pennebaker Conglomerates Inc.	Calculates the degree to which different categories of words are used and is based on scientific research connecting word use to personality and psychological states	http://www.liwc.net/
Clarabridge Platform	Clarabridge	Platform for sentiment and text analytics. Functions include information extraction, classification, and sentiment analysis	http://www.clarabridge.com/PRODUCTS/Clarabridge-Platform/
Leximancer	Leximancer	Functions include information extraction, association analysis, categorization, and visualization	https://www.leximancer.com/
Salience Engine	Lexalytics, Inc.	Text analytics engine used in social media monitoring, business intelligence, and voice of customer. Functions include information extraction, categorization, and theme recognition	http://www.lexalytics.com/technical-info/salience-engine-for-text-analysis
AlchemyAPI	AlchemyAPI, Inc.	Comprehensive text mining platform. Functions include information extraction, sentiment analysis, concept tagging, association analysis, language detection, intent mining, and topic categorization	http://www.alchemyapi.com/
VisualText	Text Analysis International, Inc.	Functions include information extraction, categorization, summarization, automatic coding, indexing, and natural language query etc.	http://www.textanalysis.com/
WordStat	Provalis Research	Quantitative text analysis tool. Functions include information extraction, categorization, clustering, association analysis, and visualization. It has seamless integration with Simstat (a statistical tool) and QDA Miner (a qualitative analysis tool)	http://provalisresearch.com/products/content-analysis-software/

Table 1 (continued)

Text miner	Vendor	Description	URL
PolyAnalyst and TextAnalysis	Megaputer	PolyAnalyst perform linguistic and semantic text analysis including coding, clustering, and taxonomy creation. TextAnalyst focuses on keyword identification and automatic summarization	http://www.megaputer.com/ site/text_mining.php
AeroText	Rocket Software	Focus on information extraction. Functions include entity identification, entity associations, event extraction, topic categorization, temporal resolution, and location resolution etc.	http://www.rocketsoftware. com/products/aerotext
KnowledgeREADER	Angoss Software	Combines text analysis and visual text discovery with sentiment analysis and predictive analytics to deliver integrated customer intelligence	http://www.angoss.com/
Semantic Annotation Server	Attensity	Text analysis is conducted based on semantic approach and through three engines including auto-classification, entity extraction, and exhaustive extraction	http://www.attensity.com/ products/technology/ semantic-server/
Rosette Linguistics Platform	Basis Technology Corp.	Suite of linguistic analysis components. Functions include automatic language identification, linguistic analysis, entity extraction, search engines, and entity translation from unstructured text	http://www.basistech.com/
VantagePoint	Search Technology, Inc.	Functions include information extraction, co-occurrence analysis, map/networks, data cleaning, user-managed Thesauri, and automated coding	http://www.thevantagepoint. com/
Oracle Text	Oracle	Uses standard SQL to index, search, and analyze texts and document. Its functions include information extraction, classification, clustering, and information visualization	http://www.oracle.com/ technetwork/database/ enterprise-edition/index-098492.html
Cogito	Expert System	Relies on semantic analysis and a semantic network to ensure a complete understanding of a text, finding hidden relationships, trends and events, and words in context	http://www.expertsystem. net/products-technology/ cogito-technology
Semaphore Content Intelligence Platform	Smartlogic Semaphore Limited	Specializes in information retrieval and text analytics. Four core modules: ontology manager, classification and text mining server, semantic enhancement server, and search application framework	http://www.smartlogic.com/ home/products/products-overview

Table 1 (continued)

Text miner	Vendor	Description	URL
Media Analysis Platform (MAP)	Sysomos Inc.	Focuses on social media. Functions include information search and extraction, key influencers identification, key conversation understanding, and theme and trend recognition etc.	http://www.sysomos.com/
Luxid	TEMIS	Functions include information extraction, association analysis, clustering, summarization, and categorization	http://www.temis.com/temis
Carrot2	Carrot Search	Open source search results clustering engine. Functions include information extraction, categorization, and theme identification etc.	http://search.carrot2.org/stable/search
HK Coder	Koichi Higuchi	Free quantitative text mining software. Functions include information extraction, co-occurrence networks, self-organizing map, multidimensional scaling, and clustering	http://khc.sourceforge.net/en/
GATE	The University of Sheffield	Open-source software. Functions include information extraction, association analysis, clustering, summarization, and categorization etc.	http://gate.ac.uk/
KNIME	KNIMEtech	Open-source data analytics, reporting and integration platform. Uses both text mining and network analysis in analyzing social media	http://tech.knime.org/knime-text-processing
RapidMiner	Rapid-I	Open-source software for text mining, data mining, and multimedia mining. Functions include information extraction, clustering, categorization, and sentiment analysis etc.	http://rapid-i.com/content/view/181/190/

product/service via both written cues and an assigned star rating. Transferring an eWOM sender's attitudes into cues contained in eWOM text is a communication process called *encoding* or *expression*. When other consumers read the eWOM on the Internet, they must first discern attitudinal information delivered in the eWOM by interpreting the star ratings and communication cues contained in the eWOM. Based on these perceived cues, eWOM readers then develop their own attitudes toward the product/service. This process of communication is called *decoding* or *impression*. Finally, eWOM readers' attitudes toward the product/service will lead to future patronage intentions.

In the eWOM communication, both cognitive and affective information is coded into a unique configuration of communication cues, and these cues are then perceived
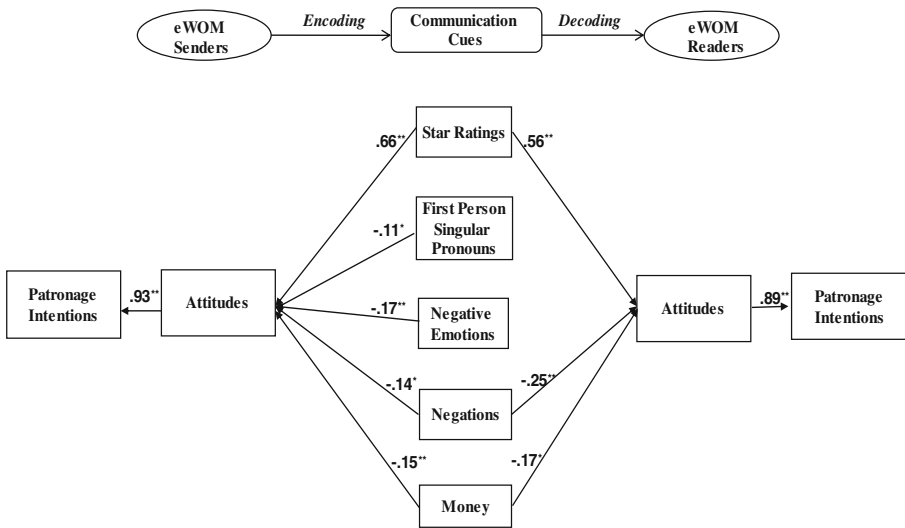
Fig. 1 Path analysis of eWOM communication

and interpreted by eWOM readers. We use attitudes to capture consumers' cognition and emotions toward a product or service. Given that eWOM is text-based communication, the cues employed by consumers in the communication are mostly linguistic cues in nature. Given the effectiveness of text mining in capturing the word uses and linguistic characteristics of text messages, we use linguistic indicators (e.g., Positive Emotions, Negative Emotions, and Pronouns) generated by text mining as the proxy for the linguistic cues employed by consumers in this study. The first objective of this study is to identify which linguistic indicators generated in text mining are significantly correlated with eWOM communicators' attitudes. In this study, we focused on the linguistic indicators that have been suggested to be significantly correlated with individuals' valenced attitudes and emotions in previous studies (e.g., Bohanek et al. 2005; Hancock et al. 2007), such as Positive Emotions, Negative Emotions, and Negations. Our first research question thus follows:

> RQ 1: Which text mining indicators are significantly correlated with eWOM communicators' self-reported attitudes?

We then tested the *predictive validity* of the linguistic indicators generated in text mining on eWOM communicators' attitudes. *Predictive validity* is the extent to which a score predicts scores on some criterion measure (Cronbach and Meehl 1955). We hereby use predictive validity to measure to what extent linguistic indicators generated by text mining predict eWOM communicators' self-reported attitudes. Specifically, we used the correlations between the linguistic indicators generated by text mining and eWOM senders' (or readers') self-reported attitudes as a measure of the validity of the linguistic indicators in predicting eWOM senders' (or readers') attitudes. Our second research question thus follows:

> RQ 2: What is the predictive validity of text mining indicators on eWOM communicators' attitudes?

Text mining is valid in eWOM research only when its indicators can precisely capture the communication cues employed by eWOM senders and perceived by eWOM readers. We therefore examine the *concurrent validity* of the text mining indicators. *Concurrent validity* is defined as the correlations between an indicator and a previously validated measure of the same or closely related construct (Alpers et al. 2005). Here, concurrent validity is used to examine to what extent text mining indicators on word uses are able to capture the word uses reported by eWOM communicators. In this study, we focus on linguistic indicators (Positive Emotions and Negative Emotions) that are associated with emotion word uses. These two indicators are measured by the percentage of positive emotion words or negative emotion words contained in the online reviews respectively. We evaluated their concurrent validity by examining the correlations between each of these two indicators and eWOM communicators' self-reported uses of either positive or negative emotion words. Our third research question thus follows:

> RQ3: What are the relationships between text mining indicators on emotion word uses and eWOM communicators' self-reported uses of emotion words?

The validity of text mining indicators was further evaluated by comparing their predictive validity to that of the star ratings on eWOM communicators' attitudes toward a product/service. Each online review contains a star rating and a text review. Serving as consumers' self-reported overall evaluation of a product/service, star ratings have been widely used in predicting consumers' attitudes, patronage behavior, and product sales in marketing research. In contrast, the text review has been largely ignored. Researchers (e.g., Kleij and Musters 2003; Owen et al. 2006) have suggested that the linguistic indicators generated in text mining and self-reports may capture different types of memories and provide distinct information about an individual. Specifically, self-reported star ratings require participants to condense their entire memory experience into few global responses (Bohanek et al. 2005), while text writing uses fine-grained, specific, and concrete information. We therefore expect that text mining indicators may capture additional information about consumers' attitudes above and beyond their self-report star ratings. Moreover, it is meaningful to introduce text mining into eWOM research only if its indicators can contribute unique explanation power of consumer attitudes above and beyond the widely used star ratings. Our fourth research question thus follows:

> RQ4: Could the text mining indicators explain additional variance in eWOM communicators' attitudes toward a product/service beyond and above the star ratings?

4 Methodology

This study used restaurants as the research context. Among the various eWOM formats (e.g., consumer forums, instant messages, and personal emails), this study focused on online reviews, as consumer-to-consumer online reviews are widely used. Two web-based self-administered surveys were employed as the predominant method of data

collection. Two independent national samples of adult consumers were purchased from a professional online survey company for conducting the two surveys respectively.

4.1 Survey procedures

The first survey focused on the encoding phase of eWOM senders. The online survey company sent emails to a random sample of its consumer panel to invite them to fill out the web-based survey. Subjects were first instructed to recall their experience with a local restaurant and then answer questions about their star ratings, emotional states, attitudes, and patronage intentions toward the restaurants. Participants were then instructed to write an online review about the restaurants and to indicate to what extent they had used positive and negative emotion words in writing their reviews. We collected 230 completed responses. Among the responses, 105 reviews that were longer than 80 words were kept. A cut-off of 80 words results from a compromise between the length recommendation of text mining software and the common length of online review practices in popular websites. A conservative recommended text length for a reliable text mining analysis is 100 words (Mehl 2010). Many popular websites, such as Amazon.com, invite their customers to write online reviews longer than 75 words.

The second survey focused on examining the decoding process of eWOM readers. Another random sample from the online survey company's panel was invited to take this survey and serves as eWOM readers. Before data collection, the 105 remaining reviews collected from the first survey were randomly divided into 35 groups with each group containing three reviews. After participants (eWOM readers) logged into the survey website, they were randomly assigned to one of the 35 groups of reviews. Participants were then instructed to imagine that they were checking the restaurant reviews on the Internet prior to their visit to a city for the first time. After reading each review, they were asked to indicate their attitudes, emotional states, and patronage intentions toward the focal restaurant as well as to what extent they had detected emotion words in reading the reviews. Reliable responses to 90 reviews were collected. These responses were then matched and merged with that of the first survey and formed the final dataset.

4.2 Measures

Star rating is a five-star scale that is widely used in popular online review forums. Attitudes were calculated by averaging the scores of its cognitive and affective dimensions. The cognitive dimension was measured by four items on a seven-point semantic differential scale adapted from Holbrook and Batra (1987). The overall positive emotions and negative emotions were measured by participants' self-reports on a seven-point Likert scale ranging from 1 (Not at all) to 7 (Extremely strong). Future patronage intentions were measured by three items on a seven-point bipolar scale adapted from Gotlieb and Sarel (1991). EWOM communicators' self-reported uses of emotion words were measured by the extent to which positive and negative emotion words were used in the reviews respectively on a seven-point Likert scale, ranging from 1 (Not at all) to 7 (Very much). Because each online review was read and rated by three to six readers in the second survey, inter-rater reliabilities were evaluated to make sure the measures of self-reported emotion word uses were reliable. Interclass correlations

show that the reliabilities for the self-reported uses of positive emotion words ($r=0.89$) and negative emotion words ($r=0.92$) were adequate.

We employed linguistic inquiry and word count (LIWC) as our text mining tool for three reasons. First, LIWC is one of the most widely text analysis tools and has been successfully used in academic research (e.g., Pennebaker et al. 2003; Tausczik and Pennebaker 2010). Second, LIWC uses word count strategy, which is the foundation of most text mining tools. It is expected that the evidence provided for the validity of LIWC can be generalized to other text mining tools. Using word count strategies, text mining software generates linguistic indicators by calculating the percentage of the counts of the words falling into a number of word categories (e.g., pronouns, emotions, and cognitions). Third, LIWC is easy to use. It automatically extracts quantitative indicators, which can be directly used in statistical analysis.

5 Data analysis and results

To answer RQ1, regarding which text mining indicators are associated with eWOM communicators' attitudes, linguistic indicators were correlated with eWOM communicators' attitudes. Results showed that Positive Emotions and its subcategories of Optimism and Energy, Negative Emotions and its subcategory of Anger, Negations, Money, Parenthesis, and Total First Person Pronouns and its subcategory of First Person Singular Pronouns were significantly associated with eWOM communicators' self-reported attitudes (see Table 2).

To answer the RQ2 regarding the predictive validity of the linguistic indicators, multiple correlations between text mining indicators and eWOM communicators' self-reported attitudes were calculated. The multiple correlations between text mining indicators and eWOM senders' and readers' self-reported attitudes are $r=0.75$ and $r=0.68$, respectively. Text mining indicators explained 56 % and 47 % variance in eWOM senders' and readers' self-reported attitudes, respectively. Above results show that text mining indicators had good predictive validity and explained a significant amount of variance in eWOM communicators' attitudes.

Table 2 Correlations between eWOM communicators' attitudes and linguistic indictors

Text mining indicators	EWOM senders' attitudes	EWOM readers' attitudes
Negations	−0.48[**]	−0.49[**]
Positive Emotions	0.36[**]	0.34[**]
Optimism and Energy	0.28[**]	0.27[**]
Negative Emotions	−0.52[**]	−0.42[**]
Anger	−0.33[**]	−0.22[**]
Money	−0.37[**]	−0.33[**]
Parenthesis	−0.26[*]	−0.22[*]
First Person Singular Pronouns	−0.28[**]	N.S.
Total First Person Singular Pronouns	−0.25[*]	N.S.

[*] $p<0.05$, [**] $p<0.01$

To answer RQ3 regarding the concurrent validity of text mining indicators, Pearson correlations were employed. Results show that the text mining indicator of Positive Emotions was positively associated with eWOM senders' self-reported use of positive emotion words ($r=0.31$, $p<0.01$) and with eWOM readers' perceived use of positive emotion words by eWOM senders ($r=0.32$, $p<0.01$). The text mining indicator of Negative Emotions was positively associated with eWOM senders' self-reported use of negative emotion words ($r=0.54$, $p<0.001$) and with eWOM readers' perceived use of negative emotion words by eWOM senders ($r=0.44$, $p<0.001$). These moderate to high correlations supported the concurrent validity of text mining indicators.

To answer the last research question regarding the relative predictive validity of text mining indicators to the star ratings, a path analysis was conducted (see Fig. 1). Those text mining indicators that were significantly correlated with eWOM communicators' attitudes were incorporated into the model as the predictors. All variables were treated as manifest variables and Lisrel 8.8 was employed to conduct the analysis. Initial results revealed that text mining indicators were redundant. After dropping the insignificant paths, only the star ratings and text mining indicators of Negative Emotions, Negations, Money, and First Person Singular Pronouns were retained in the model. The overall fit of the final model was acceptable, $\chi^2_{(16)}=48.63$, NFI=0.94, NNFI=0.90, CFI=0.96, RMSEA=0.13. The star ratings and text mining indicators together explained 80 % and 57 % of the variances in eWOM senders' and readers' attitudes, respectively. The model explained 86 % and 79 % of the variance in eWOM senders' and readers' future patronage intentions, respectively. The star ratings were the strongest predictor of eWOM communicators' attitudes and alone explained 72 % and 49 % of the variance in eWOM senders' and readers' attitudes, respectively. Text mining indicators explained 8 % additional variance in both eWOM senders' and readers' attitudes above and beyond the star ratings.

6 Discussion

This study identified the text mining indicators (e.g., Negative Emotions and Negations) that can significantly predict eWOM communicators' attitudes. The result is largely consistent with the findings in social psychology (e.g., Hancock et al. 2007) which suggested that the uses of some linguistic cues are significantly associated with individuals' emotional states and attitudes. Beyond the previous studies, we found that the word count of Money, which contains the words related to price (e.g., price, cheap), is negatively associated with eWOM communicators' attitudes. This result is not surprising given that high price or low value is a major source of customer dissatisfaction. We also found that Parentheses are negatively associated with eWOM communicators' attitudes, which has never been reported in previous studies. By checking the original online reviews, we found that some consumers used parenthetical comments to provide additional explanations of their negative experience. We also found that text mining indicators are better in predicting eWOM senders' than eWOM readers' attitudes. Specifically, they have stronger multiple correlations with eWOM senders' attitudes than with eWOM readers' ($r=0.75$ vs. $r=0.68$). Star ratings and text mining indicators together explained more variance in eWOM senders' than in eWOM readers' attitudes (80 % vs. 57 %).

Some systematic differences exist between eWOM senders and readers regarding their configurations of linguistic cues when encoding versus decoding eWOM information. For example, four text mining indicators (including First Person Singular Pronouns, Negative Emotions, Negations, and Money) made unique contributions to eWOM senders' attitudes above and beyond the star ratings. In contrast, only two text mining indicators (including Negations and Money) contributed uniquely to eWOM readers' attitudes. Aside from the above differences, the configurations of linguistic cues employed by eWOM senders and readers are remarkably consistent. In summary, eWOM readers' decoding system is largely consistent with eWOM senders' encoding system, but they are not perfectly matched. This makes eWOM communication effective in general but with some possibility of miscommunication.

We found that the correlations between text mining indicators of emotion word use and eWOM communicators' self-reported use of emotion words were significant but not very strong ($r<0.6$). One possible explanation is that the word categories employed by consumers may not exactly correspond to the word counts calculated in text mining software (Alpers et al. 2005). For example, the word category of Negations which includes words such as "never" and "not" and represents a denial, contradiction, or negative statement is very likely considered by consumers as negative emotion words as supported by the strong correlations between Negations and eWOM communicators' self-reported use of negative emotion words ($r=0.55$, $p<0.01$).

Our results also show that text mining indicators have the potential to provide a valid supplemental measure of the star ratings. In eWOM research, star ratings have been the most widely used proxy of eWOM valence. Consistently, we found that star ratings are the most powerful predictors of eWOW communicators' attitudes. On the other hand, text mining indicators also have good predictive utilities in eWOM communicators' attitudes and can explain 8 % of the additional variance in them above and beyond the star ratings. All the indicators that contribute uniquely to eWOM communicators' attitudes are negative-valenced cues. These results can be explained by the accessibility–diagnosticity model (Feldman and Lynch 1988; Herr et al. 1991). In online reviews, star ratings are the most accessible information because they always stand out at the top of reviews, making them the most important eWOM indicators. Negative information is more thought-provoking and perceived to be more diagnostic than positive or neutral cues (Ahluwalia 2002). Thus, negative-valenced linguistic indicators are very likely to be considered as an input in communicating and developing consumers' attitudes.

7 Implications

This paper offers initial evidence for the validity and utility of text mining in studying eWOM communication and provides significant implications for marketing researchers and practitioners. Text-based online communication has become an inseparable part of modern consumers' daily lives and contains valuable market information. However, the current eWOM research relies heavily on the star ratings while the text content of eWOM has been largely neglected. Our results show that by utilizing text mining, researchers and practitioners may easily capture consumers' attitudes toward a certain product/service. Moreover, the text mining can explain additional variance in consumers' attitudes above and beyond the star ratings. Thus, the text mining indicators

may provide a promising supplement to the widely used star ratings as an indicator of eWOM valence. Using both star ratings and text mining indicators may generate a fuller picture of the consumer market. In addition, the text mining indicators can be easily used in different statistical analysis such as categorization, trend identification, and association analysis to provide insights about the market, competitors, and consumers.

8 Limitations and future research

Text mining, as a relatively new method, has some limitations (Mehl 2006). First, we found that the word categories employed in text mining programs do not accurately catch the words used by consumers. Thus, the dictionary in text mining software should be adapted to the consumer research context. Second, most text mining techniques are based on the word count strategies and cannot provide a reliable and valid result when the text is short. For example, the percentage of positive emotion words used is commonly used to reflect the positive emotion contained in the text. In a short online review like "meals taste great, great service! Too expensive!", because it has two positive emotion words ("great," "great") out of seven words in total, the percentage of positive emotion words will be $2/7=28.6$ %. However, the percentage of positive emotion words is much lower in regular online reviews. For example, the average percentage in the sample used in this study is 4 %. Thus, text mining results for short texts might be highly biased. Third, this study employed LIWC, one popular text analysis tool. There are many other text mining tools in the market. Studies should be conducted to evaluate and compare the advantages and disadvantages of these text mining tools. Finally, this study focused on online reviews. Further research efforts should be extended to examine the validity and utility of text mining in studying other text-based communications, such as Facebook, Twitter, complaint letters, and emails.

References

Aggarwal, P., Vaidyanathan, R., & Venkatesh, A. (2009). Using lexical semantic analysis to derive online brand positions: An application to retail marketing research. *Journal of Retailing, 85*(2), 145–158.

Ahluwalia, R. (2002). How prevalent is the negativity effect in consumer environments? *Journal of Consumer Research, 29*, 270–279.

Alpers, G., Winzelberg, A., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C. (2005). Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior, 21*(2), 361–376.

Bantum, E., & Owen, J. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment, 21*(1), 79–88.

Bohanek, J., Fivush, R., & Walker, E. (2005). Memories of positive and negative emotional events. *Applied Cognitive Psychology, 19*(1), 51–66.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity for psychological tests. *Psychological Bulletin, 52*, 281–302.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73*(3), 421–435.

Gotlieb, J., & Sarel, D. (1991). Comparative advertising effectiveness: The role of involvement and source credibility. *Journal of Advertising, 20*(1), 38–45.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60–76.

Hancock, J., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. In *CHI 2007 proceedings of the SIGCHI conference on human factors in computing systems* (pp. 929–932).

Henning-Thurau, T., Gwinner, K., Walsh, G., & Gremler, D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing, 18*(1), 38–51.

Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility–diagnosticity perspective. *Journal of Consumer Research, 17*(March), 454–462.

Holbrook, M., & Batra, R. (1987). Assessing the role of emotions as mediators of consumer responses to advertising. *Journal of Consumer Research, 14*(3), 404–420.

Jansen, J. (2010, September 29). Attention shopper: Online product research. Pew Internet and American Life Project, available at www.pewinternet.org/pubs/1747/e-shopping-researched-product-service.aspx (Accessed 1/25/2012).

Kahn, J., Tobin, R., Massey, A., & Anderson, J. (2007). Measuring emotional expression with the linguistic inquiry and word count. *American Journal of Psychology, 120*(2), 263–286.

Kleij, F., & Musters, P. (2003). Text analysis of open-ended survey responses: A complementary method to preference mapping. *Food Quality and Preference, 14*(1), 43–52.

Kozinets, R. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research, 39*(1), 61–72.

Lehto, X., Park, J., Park, O., & Lehto, M. (2007). Text analysis of consumer reviews: The case of virtual travel firms. In M.J. Smith, G. Salvendy (Eds.), *Human Interface, Part I, HCII 2007, LNCS4557*, 490–99.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing, 70*(3), 74–89.

Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 141–156). Washington, DC: American Psychology Association.

Mehl, M. R. (2010). Automatic text analysis. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for behavioral research on the Internet* (pp. 109–127). Washington, DC: American Psychology Association.

Owen, J., Giese-Davis, J., Cordova, M., Kronenwetter, C., Golant, M., & Spiegel, D. (2006). Self-report and linguistic indicators of emotional expression in narratives as predictors of adjustment to cancer. *Journal of Behavioral Medicine, 29*(4), 335–345.

Pennebaker, J., Mayne, T., & Francis, M. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72*(4), 863–871.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language: Our words, our selves. *Annual Review of Psychology, 54*(1), 547–577.

Ramanathan, V., & Meyyappan, T. (2013). Survey of text mining. In *Proceedings of International Conference on Technology and Business Management* (pp. 508–514).

Sirdhar, S., & Srinivasan, R. S. (2012). Social influence effects in online product ratings. *Journal of Marketing, 76*(5), 70–88.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24–54.