Journal of Marketing Analytics (2023) 11:662–676
https://doi.org/10.1057/s41270-023-00218-6

ORIGINAL ARTICLE

Extracting marketing information from product reviews: a comparative study of latent semantic analysis and probabilistic latent semantic analysis

Shimi Naurin Ahmad[1] · **Michel Laroche**[2]

Revised: 13 January 2023 / Accepted: 6 March 2023 / Published online: 8 April 2023

Abstract
User-generated content (UGC) contains customer opinions which can be used to hear the voice of customers. This information can be useful in market surveillance, digital innovation, or brand improvisation. Automated text mining techniques are being used to understand these data. This study focuses on comparing two common text mining techniques namely: Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (pLSA) and evaluates the suitability of the methods in two differing marketing contexts: Reviews from a product category and from a single brand from Amazon. The objectives of review summarization are fundamentally different in these two scenarios. The first scenario can be considered as market surveillance where important aspects of the product category are to be monitored by a particular seller. The second scenario examines a single product, and it is used to monitor in-depth customer opinions of the product. The results support that depending on the objective, the suitability of the technique differs. Different techniques provide different levels of precision and understanding of the content. The power of machine learning methods, domain knowledge and Marketing objective need to come together to fully leverage the strength of this huge user-generated textual data for improving marketing performance.

Keywords User genrated content (UGC) · Big data · LSA · pLSA · Text mining · Amazon reviews

Introduction

Due to the proliferation of the use of internet, e-commerce and social media, text data are readily available on the web. These text data can be a great resource for marketers who are eager to listen to the customers to better manage the marketing process. Traditionally, marketers send surveys to the customers; nowadays, product reviews, blogs or other digitized communication provide information on the attributes that are relevant to marketing decision making such as, adoption of a new product, possible composition of consideration set, etc. (Lee and Bradlow 2011). Due to the large volume and unstructured nature of the text data, sophisticated modeling is warranted, and researchers have been using data analytics, specifically machine learning techniques to uncover various patterns that helps the business (Mikalef et al. 2020b).

A large number of studies have examined the effect of big data analytics on firm's performance and the evidence quite overwhelmingly suggests that data analytics improves firm's decision making and innovation (Branda et al. 2018; Gupta and George 2016), customer relationship management, management of operations risk and efficiency, market performance (Wamba et al. 2017) and at the end, overall performance (Kiron 2013). By providing accessible information to managers, data analytics creates a competitive advantage (Mikalef 2020a). Studies have also shown a positive association between customer analytics and firm's performance (German et al. 2014). Customer analytics may tap into different areas of customer experience ranging from purchasing behavior, prediction of buying trend to product recommendation, co-creation (Acharya et al. 2018) and opinion summarization about a specific feature of a product, etc.

✉ Shimi Naurin Ahmad
shimi.ahmad@morgan.edu

Michel Laroche
michel.laroche@concordia.ca

[1] Department of Business Administration, Earl G. Graves School of Business and Management, Morgan State University, 1700 East Cold Spring Lane, Baltimore, MD 21251, USA

[2] Department of Marketing, John Molson School of Business, Concordia University, 1450 Rue Guy, Montréal, Québec H3G 1M8, Canada

The task of summarizing opinions or reviews has become one of the central research areas among the text mining community, mainly in the information retrieval literature (Mudasir et al. 2020; Hu et al. 2017). The techniques are becoming more sophisticated, and studies are increasingly reporting methods for extracting aspects/topics, textual summaries, etc. (Mudasir et al. 2020). The different formats and techniques provide different levels of understanding or precision of the content. Therefore, the users need to adapt these methods according to their own needs.

In the influential paper on automated text analysis in marketing, Berger et al. (2020) emphasized that regardless of the focus (to make a prediction, to assess impact or to understand a phenomenon), "doing text analysis well requires integrating skills, techniques, and substantive knowledge from different areas of marketing." (Berger et al. 2020, p. 6). Text analysis yields its best result when the positivist analysis (the factual knowledge gained by a scientific process, usually a quantitative method) is used in combination with qualitative and interpretive analysis. For example, Kubler et al. (2017) used tailored marketing dictionary which allows the analysis to be interpreted in the marketing context, rather than in a general context. A word may have a different interpretation depending on the context where it is used. The author then utilized this exclusive dictionary (tailored for marketing domain) inside an automatic text analysis-based sentiment extraction tool, namely, support vector machine (Cui and Curry 2005) to uncover different marketing metrics from user-generated content. Berger et al. (2020) further elaborated this point and explained that quantitative skill helps building the right mathematical model, but behavioral skill relates the phenomenon (the findings) to underlying psychological processes, and most importantly for marketers, strategy skill which can be defined as the skill to understand the findings from the big data and convert these findings into firm's actionable items and outcomes helps reach firm's goals. Therefore, these text data can ultimately aid a firm's marketing decision making and be a great resource, but the combinations of above mentioned tools seem to be necessary. In this light, it is very important that marketers build their tools using machine learning techniques as well as the other soft skills, especially Marketing-specific knowledge and skill to get most out of the data (Ma and Sun 2020). However, Marketing analytics literature is still premature in providing guidance about the suitability of a particular analytics tool in crafting overall firm's strategy (Vollrath and Villegas 2022). Machine learning and text mining experts are skilled in building accurate and precise mathematical models and often, their goal is to improve prediction. The "right answer" for goal might be different for different objectives. Therefore, when the goal is to improve overall marketing metrics, it is recommended in the literature that domain knowledge be incorporated in the process (Hair and Sarstedt 2021). In a recent paper, Huang and Rust (2021) elaborated that Artificial Intelligence use in Marketing should be in three stages: "Mechanical AI" for repetitive tasks, "Thinking AI" for analyzing data and making a decision and "Feeling AI" for understanding consumers and interacting with them. The latter two need domain knowledge as input to optimize the goal of improving marketing metrices. This paper responds to that call of integrating quantitative model with goal-specific domain knowledge to better assist managers in taking actions. A firm's marketing decision making through text analysis task is better served when domain knowledge is incorporated rather than borrowing predefined model invented for a different purpose.

As mentioned before, opinion summarization has been an active area of research in information retrieval literature for over decades now, marketers need to tailor these methods according to their objectives and needs to leverage the strength of this huge textual data. The strength of statistical power and goal of marketers need to come together to fully utilize this opportunity. With this in mind, the current research focuses on comparing two common text mining techniques from a marketer's perspective. Analyzing the text data of the reviews posted on Amazon.com, the current study compares: Latent Semantic Analysis (LSA) (Deerwester et al. 1990) and Probabilistic Latent Semantic Analysis (PLSA) in extracting useful summarizing information in terms of common themes. In the first context, the reviews are taken from the category of kitchen appliances where there were different brands and several kitchen products within this dataset of reviews. Second, only one brand of a product's review is examined in handbag category. The objectives of review summarization are fundamentally different in these two scenarios when the analysis is intended to provide information about market research. The first scenario provides information about the whole market in that product category. It can be considered as market surveillance where important aspects of the product category are to be monitored by a particular seller to find out what characteristics of the product category are of main concern. These are also the key aspects of the whole customer experience that determine customer satisfaction or dissatisfaction. The insight can be used to improve an offering through innovation or by combining digital aspect to it, also known as digital innovation (Sahut et al. 2020). Since the information comes from well-represented consumers are "organic' text, it is free from any bias and doesn't restrict any topic which is a common problem even in well-crafted surveys (Savage and Burrows 2009). The second scenario examines a single brand. This is useful for brand managers when an in-depth analysis of consumers' opinion is sought after. There have been studies that have looked at the performances of LSA and PLSA (Ke and Luo 2015; Kim and Lee, 2020). However, to the best of our knowledge, there is no study that compares these two

methods in two different scenarios where marketing goals are different and evaluate the suitability of the techniques in differing contexts.

The rest of the paper is organized as follows: We review the literature on User-generated content, Customer analytics, opinion summarization and some text mining tools. Experimentation is presented next along with findings, followed by the Discussion and managerial implications.

Literature review

User-generated content

UGC refers to any content created by users or consumers of a product or service, such as product reviews, social media posts, blog articles, and videos. In recent years, user-generated content (UGC) has exploded, and these UGCs are often text data in the form of blogs, reviews, or social media interactions. The scholars have examined a range of issues (Iacobucci et al. 2019), such as how and why people make UGC contributions (Braune and Dana 2021; Moe and Schweidel 2012; Ransbotham et al. 2012) and the impacts of UGC (Zhang et al. 2012) including review rating and text (Sallberg et al. 2022), among others.

User-generated content (UGC) can benefit firms in several ways, including increased customer engagement (Bijmolt et al. 2010), improved brand loyalty (Llopis-Amorós et al. 2019), and brand co-creation (Koivisto and Mattila 2020). A study by Constantinides and Fountain (2008) found that UGC can positively impact the credibility and perceived quality of a brand, leading to increased brand loyalty and purchase intentions. Additionally, UGC can enhance the authenticity of a brand by providing real-life examples of product usage and customer experiences. More importantly, UGC can also provide valuable insights into customer preferences, needs, and touch points, which can help firms improve their products and services. In a study by Bernoff and Li (2008), it was found that UGC can help firms identify customer needs and trends, leading to improved innovation and product development.

In a study of UGC and its impact, Li et al., (2021) modeled consumer purchase decision process and found evidence that UGC impacts every state of this process. UGC can also provide valuable insights and ideas that firms can use to develop new products, services, or marketing strategies (Hanna et al. 2011).

Although UGC can be generated in various forms, product reviews ratings and content are very influential in terms of sales (Mudambi and Schuff 2010). The impact of review ratings on product sales has been thoroughly studied (Chevalier & Mayzlin 2006; Liu 2006) including various product categories. The sales of books (Chevalier & Mayzlin 2006) and movies (Liu 2006) were affected by ratings of the review generated by users. Research has also explored the impact of review content on marketing parameters, such as the helpfulness vote (Ghose and Ipeirotis 2011), consumer engagement (Yang et al. 2019) and digital innovation (Sahut et al. 2020). Although these data can provide valuable information about market and customers, it can be hard to decipher the actual information from the unstructured data (Zhu et al. 2013) and gave rise to customer analytics.

Customer analytics

As mentioned before, a large number of studies investigated the relationship between big data analytics or customer analytics and results suggest that data analytics enhances firm's decision making and innovation (Branda et al. 2018; Gupta and George 2016). To analyze the customer generated text data, which most commonly occur across web, marketing scholars are using text analysis tools and methods to analyze these data automatically (Kamal 2015). These data types and analytical methods vary widely across different branches of Marketing analytics (Iacobucci et al. 2019). There are many cutting edge methods that have been used by Marketing scholars to analyze UGC and consumer reviews, in particular.

Ghose and Ipeirotis 2011 showed strong evidence that consumer review affect economic outcome, product sales and some aspects of reviews such as subjectivity, informativeness, readability, and linguistic correctness in reviews affects potential sales and perceived usefulness. They use Random forest model and text mining to uncover the insight. Netzer et al. 2012 came up with a market-structure perceptual map using consumer review data on diabetes drugs and sedan cars. The authors utilized the combination of text mining techniques and network analysis to introduce this map.

With a little bit different focus, Hou et al. (2022) studied driving factors of web-platform switching behavior using dataset of both blogging and microblogging activities of the same set of users. The authors used a sophisticated text analysis technique: multistate survival analysis. Skeen et al. (2022) took a very innovative approach to combine qualitative analysis with natural language processing and designed a mobile health app which was very customer centered.

Given this huge amount of user-generated content, it is quite useful to summarize consumers' opinion in the aggregate level and derive marketing information from there. Li and Li (2013) summarized a large volume of microblogs to discover Market intelligence. Since our study is closely related to this area, we next review the literature on opinion summarization and sentiment analysis.

Opinion summarization and sentiment analysis in marketing

As the name implies, opinion summarization provides an idea about the whole document collection in brief. There is vast research investigating algorithms for summarization using different technical methods (Moussa et al. 2018). In Marketing related opinion summarization techniques, Vorvorean et al. (2013) introduced a method of using social media analytics that can decipher the topics of UGC, assess a major event and at the end, can have useful impact on marketing campaign.

Sentiment classification is one of the important steps in analyzing text data and can be used as part of opinion summarization. In this process, orientation of sentences or the whole documents are identified. This will result in an overall summarization of the documents as users get an idea about what is being said (positive and negative). There are several approaches in identifying sentiments which find out the adjective in the text and thus try to understand the positivity or negativity of the text (Li et al. 2018; Salehan and Kim 2016). Salehan and Kim (2016) used sentiment analysis to see the impact of online consumer review in terms of their readership and helpfulness.

Sentiment classification can be used as a simple summary, this method is very useful when there is a large collection of data involved and aggregate level opinion is sought after. Some technical methods studies (Jimenez et al. 2019; Kamps and Marx 2001) used WordNet-based approach using semantic distance from a word to "positive" and "negative" as a classification criterion between sentiments. Ku et al., (2006) used frequency of the terms for feature identification and used sentiment words to assign opinion scores. Lu et al. (2009) used natural language processing techniques to K (K = any number) interesting aspects and utilized bays classifier for sentiment prediction.

As mentioned before, extracting common themes along with its sentiment from user-generated content can be considered as summarizing the content since it tends to reflect the whole content. Next, we review some of the text analysis techniques that have been used in prior research.

Text analysis tools and methods

Studies have used a wide variety of techniques to analyze texts and specially to extract themes from texts. One of the foundational techniques to extract themes from a body of text is Latent Semantic Analysis (LSA). There are many studies that used LSA for the purpose of opinion summarization (Steinberger and Ježek 2009). Sidorova et al., (2008) used LSA to uncover the intellectual core of information research from published journal papers. The method mainly relies on the co-occurrence of the word and is not based on statistical modeling. Cosine distance can also be used in latent semantic analysis space to measure topics in the text (Turney and Littman 2003).

Another stream of techniques that focuses on extracting themes is defined as generative probabilistic model and is based on a solid foundation of statistics. Vocabulary distribution is used to find topics of texts. Basically, it first identifies the word frequencies and relation between other words (co-occurrences) effectively. There are several topic modeling approaches in this family. Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999) and LDA (Latent Dirichlet Allocation) are the important ones. Table (Table 1) shows that identifies some key literature using these methods:

Comparative studies between LSA and PLSA

There are some studies that have compared these two techniques (LSA vs. PLSA) in various contexts. One study (Kim et al. 2020) compared two text mining techniques to predict blockchain trends by analyzing 231 abstracts of papers and their topics. The techniques were W2V-LSA which is an improvised version of LSA and PLSA. The study concluded that the new technique W2V-LSA worked better in finding out proper topics and in showing a trend. Ke and Luo (2015) compared LSA and PLSA as automated essay scoring tools.

Table 1 Text analysis tools in prior research

	Purpose of the Study	Topic extraction approach
Sidorova et al. (2008)	To summarize journal papers of information research	LSA
Ansari et al. (2018)	Movie Recommender system	Novel Topic Modeling Technique
Zhong and Schweidel (2019)	Detecting shift of UGC	Variation of LDA
Timoshenko and Hauser (2019)	To assess the usefulness of automatic techniques to identify consumer needs	Convolutional Neural Network
Liu et al. (2019)	How consumers use review content	Deep learning for natural language processing
Yu et al. (2012)	Predicting sales of movies	PLSA

The result showed that both methods have some correlation in their performances, and both did well in their task. A bit different, a study by Cvitanic et al. (2016) compared the suitability of using LDA and LSA in the context of textual content of patents. The study suggested that more work is needed to recommend one method versus another to analyze and categorize patents.

Although along the same line, the current study does *not* fully focus on summary presentation; instead, it focuses on features and their sentiment orientation that are visible in the topics. Summary presentation is often used to make the summary of the reviews more understandable to customers. From a managerial perspective, they need to know in detail what is being said about a particular feature. Therefore, the current study examines the topic extraction and the suitability of these two techniques from a managerial perspective. As mentioned in the previous paragraph, there have been studies where performance of these two techniques is compared. Some of them found evidence of the superiority of one method, some reported the same kind of efficiency, and some recommended more studies to conclude. However, to the best of our knowledge, no study has looked at these methods in two different contexts with varying objectives. Given the new understanding of automatic text analysis, where quantitative skill is to be combined with domain knowledge, and the fact that accuracy of retrieval is not the focus in marketing, the current study tries to fill the void in research in this area.

Methods and data

For the purpose of this study, as a starting point of domain-specific tool adaptation, we use two fundamental techniques (LSA and PLSA) of topic modeling. Both use topic modeling algorithms and the basic assumption of this type of modelling algorithms are *(a)* each document consists of a mixture of topics, and *(b)* each topic consists of a collection of words. LSA is one of the foundational techniques in topic modeling. LSA takes a document and terms matrix and decompose it in two reduced dimension matrices: one is document-topic matrix and the other is topic-term matrix. The whole technique is based upon singular value decomposition (SVD) and dimension reduction. pLSA, on the other hand, belongs to another stream of techniques within topic modeling. It is based on probabilistic method; Instead of SVD used in LSA, pLSA tries to come up with a probabilistic model with latent topics which can ultimately reproduce the data. There are other topic modeling techniques that build on pLSA like LDA (Latent Dirichlet Allocation) which is basically a Bayesian version of pLSA and therefore uses Dirichlet priors. Next, we describe the methods in detail:

Latent semantic analysis

Latent Semantic Analysis (LSA) is a text mining technique that extracts concepts hidden in text data. This is based solely on word usage within the documents and does not use a priori model. The goal is to represent the terms and documents with fewer dimensions in a new vector space (Han and Kamber 2006). Mathematically, it is done by applying singular value decomposition (SVD) on a term-by-document matrix (X) that holds the frequency of terms in all the documents of a given collection. When the new vector space is created by retaining a small number of significant factors k and X is approximated by $X = T_k S_k D_k^T$ (Landauer et al. 1998). Term loadings ($L_T = T_k S_k$) are rotated (varimax rotation is used) to obtain meaningful concepts of the document collection. The algorithm is shown in Fig. 1. It is implemented using Matlab.

Probabilistic latent semantic analysis

Probabilistic Latent Semantic Analysis (pLSA) is another text mining method that was developed after LSA (Hofmann 1999). Unlike LSA, it is based on a probabilistic method, namely, a maximum likelihood model instead of a Singular Value Decomposition. The goal is to recreate the data in terms of term-document matrix by finding the latent topics. So, a model P(d,w) is put forward where document d and word w are in the corpus and P(d,w) corresponds to that entry in the document-term matrix. In this scenario, a document is sampled first, and in that document, a topic z is sampled, and based on the topic z, a word w is chosen. Therefore, d and w are conditionally independent given a hidden topic 'z'. This can be represented in Fig. 2:

A document can be selected from the corpus with a probability of P(d). In the selected document, a topic z can be chosen from a conditional distribution with a probability P(z|d) and a word can be selected with a probability of P(w|z). The model makes two assumptions. First, the joint variable (d,w) is sampled independently, and more importantly, words and the documents are conditionally independent.

$$P(d, w) = P(d)P(w|d)$$

After some mathematical manipulation, it can be written in the following form.

$$P(d, w) = P(d) \sum P(z|d)P(w|d)$$

The modeled parameters are commonly trained using an Expectation–Maximization algorithm. The equation lets us estimate the odds to find a certain word within a chosen document using the likelihood of observing some document and then based upon the distribution of topics in that

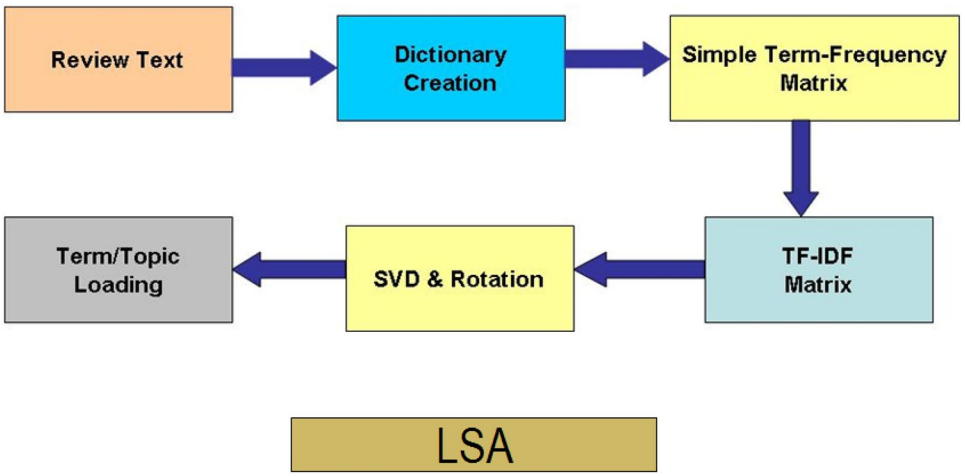

Fig. 1 Algorithm flow chart (LSA)

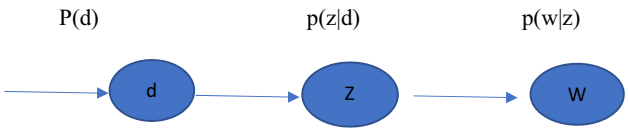

Fig. 2 PLSA model

document, the odds to find a certain word within that topic. In a flowchart form (Fig. 3):

Differences between LSA and PLSA

Both LSA and PLSA can recreate the data content based on the model. But there is an important difference between the two methods.

First, in LSA calculations, SVD is based on Matrix decomposition which is the F-norm approximation of the term frequency matrix, while PLSA relies on the likelihood function and prior probability of the latent class (probability of seeing this class in the data for a randomly chosen record, ignoring all attribute values) and, finds the maximum conditional probability of the model.

Second, in LSA, the recreated matrix X does not contain any normalized probability distribution, while in PLSA, the matrix of the co-occurrence table is a well-defined probability distribution. Both LSA and PLSA perform dimensionality reduction: LSA keeps only K singular values and PLSA, keeps K aspects.

For the purpose of the comparison, in the subsequent sections we need to find the comparable parameters of both models. From the mathematical and interpretation standpoint, the three matrices from SVD correspond to three probability distributions of PLSA:

(a) T Matrix is synonymous to P(d|z) (doc to aspect).
(b) D Matrix is related to P(z|w) (aspect to term).
(c) S Matrix related to P(z) (aspect strength).

Performance Measure

To compare two techniques, one needs to evaluate the performance of each of these methods. In the analysis section, both quantitative evaluation and qualitative observations (Mei et al. 2007; Titov and McDonald 2008) are used to analyze the data results. Among the quantitative measure, precision/recall curve is the most widely used measure (Titov and McDonald 2008). Precision is defined as the number of relevant words retrieved divided by number of all words retrieved. This provides a measure of accuracy. The numbers of irrelevant words are counted to evaluate lack of accuracy.

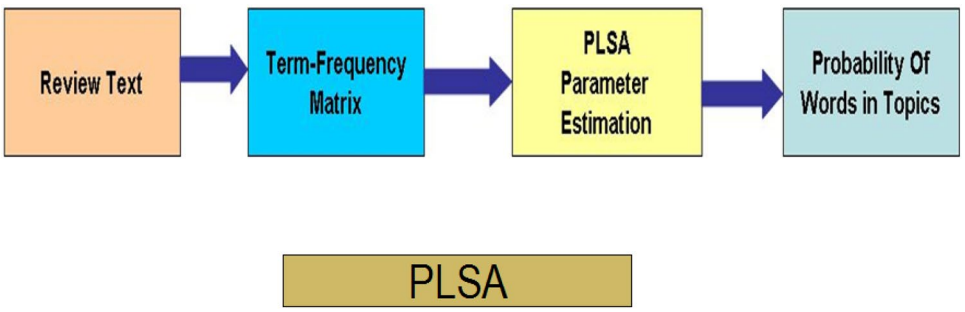

Fig. 3 Algorithm flow chart (PLSA)

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

Moreover, the following classification helps in the measure of accuracy:

	Relevant	Nonrelevant
Retrieved	True positives (tp)	False positives (fp)
Not retrieved	False negatives (fn)	True negatives (tn)

Here, we measured the false positives and compared the two techniques. Ideally, false positives should be as low as possible. The measure of recall is used when the total of relevant words is known. Since, for conversational text, it is difficult to develop and measure a list of total relevant words, we did not use recall or false positive/negative as a measure of performance in this analysis.

Data

To begin, we utilized a dataset containing reviews of kitchen appliances. It was sourced (downloaded) from publicly available dataset collected by Blitzer et al. (2007). There were also reviews on books in this dataset. We excluded book reviews, because the content of the book written in the review may confound the topics of the review. In total, there were 406 kitchen appliances reviews included in the dataset, with 148 reviews being positive and 258 reviews being negative. Additionally, the authors analyzed a second dataset consisting of reviews for a specific brand of handbag, "Rose Handbag by FASH," that was obtained from Amazon.com in 2011. This dataset contained a total of 389 reviews. We used LSA and pLSA to extract hidden topics and associated words from both datasets, and subsequently compared the performance accuracies of the two methods.

Results

First, we analyze the brand-specific Handbag reviews. The reviews which got star rating 3 or more were classified in the positive reviews. On the other hand, reviews with star ratings 1 and 2 are classified as negative reviews. In the LSA model, three dimensions are retained after SVD. To compare the extracted topics with the topics extracted from the PLSA, we kept three topic groups for PLSA too (dimensions in LSA are comparable to topics in PLSA, shown in Table 2). For the positive reviews, the three topics/factors are named as "Leading positive attributes of the product", "Core functionalities", and "Affective" based on the associated words retrieved by both methods. On the other hand, for the negative reviews, the three topics are "not leather', "Problems", "Service failure" (shown in Table 3).

The comparison of the word associated with each positive topic (Table 2) shows that topics extracted by PLSA have more interpretability and contain more information. For example, for the positive reviews, the words which have high probability to be in the topic ("Leading Positive Attribute of the Product") are "large", "roomy", "price", "quality" (colored in pink). However, these important terms (since these words imply the competitive advantage of the brand and the topic) were not picked up by LSA. Moreover, among the words picked up by LSA, "review", "purse", "thank", "shoulder" (colored in orange) is not relevant to this topic. The remaining words both in LSA and PLSA (colored black) contribute to the meaning of the factors (in both LSA and PLSA they are either relevant or neutral words). By neutral, we mean the words that are relevant and contribute to the better interpretation of the factor, but do not have unique power like the pink words in PLSA. For example, "amazing", "beautiful", "nice", etc. contribute to the meaning of the "leading positive attributes" and help in the interpretation that customers are happy with these attributes of the product. However, these do not describe any of the leading attributes. The results show the top 10 terms (according to the probability for PLSA and loadings for LSA). A

Table 2 Comparison of PLSA and LSA factors (and associated words) of the positive reviews of handbag

Factors/Topics	PLSA	LSA
Leading positive attributes of the product	Large, Roomy, Stylish, Price, Quality, Amazing, Beautiful, favorite, Bag, outfit	Beautiful, Nice, Color, Design, Happy, Thank, shoulder, picture, review, purse
Core Functionalities	Shoulder, Strap, Texture, Material, Pattern, Double, Zipper, Pocket, inside, fashion	Shoulder, Strap, Pattern, pocket, Zipper, inside, price, pretty, color, order
Affective	Birthday, Gift, Friend, Love, Pretty, Sister, Happy, Pink, Picture, Look	Birthday, Gift, Fun, Sister, Love, Favorite, Happy, Price, Absolute, please

Table 3 Comparison of PLSA and LSA factors (and associated words) of the negative reviews of handbag

Factors/Topics	PLSA	LSA
Not Leather	Plastic, Leather, Real, Expect, Zip, Spacious, Pink, Bad, Bag, color,	Plastic, Leather, Pleather, Real, Expect, Bad, Boo, Spacious, Bag, zip
Problems	Color, material, Look, Photo, Picture, Rough, Thread, Handbag, Leather, Pleather,	Deceive, Pink, Picture, Peach, Issue, color , Seller, Ugly, Massive, Photo
Service Failure	Broken, pieces, contact, Customer, Help, Product, Quality, Attach, phone, Amazon	Break, Pieces, Cheap, Faulty, phone, Receive, Decent, Zip, Close, Money

Table 4 Positive reviews relevant words extracted by both methods

	PLSA (Retrieved Relevant Words)	LSA (retrieved relevant words)
Leading positive attributes	Large, roomy, stylish, price, quality	Color, design
Core functionalities	Shoulder, strap, texture, material, double, zipper, pocket, inside, pattern	Shoulder, strap, pocket, inside, zipper, pattern
Affective	Birthday, gift, friend, love, sister, happy	Birthday, gift, fun, sister, love, happy

Table 5 Positive reviews irrelevant words extracted by both methods

	PLSA (retrieved irrelevant words)	LSA (retrieved irrelevant words)
Leading positive Attributes	Outfit	Thank, picture, review, purse
Core functionalities	Fashion	Order, price, pretty, color,
Affective		Price, absolute, please

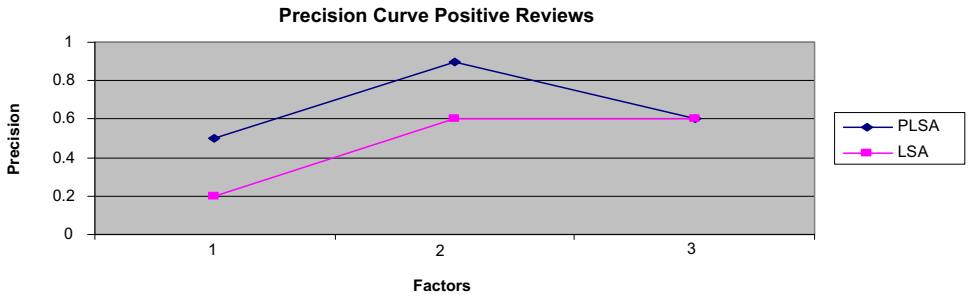

Fig. 4 Precision curve of positive reviews

comparison of relevant and irrelevant words picked up by both methods are presented in Tables 4 and 5, respectively.

To quantify the performance superiority of one method over the other, precisions of the two methods are calculated and shown graphically in Figs. 4 and 5. The number of irrelevant words picked up by both methods implies the inferiority of the method. This is shown in Table 5. A method needs to yield a high precision as well as low irrelevant words to be considered as superior technique. As mentioned before, there are some words that are neutral: neither uniquely relevant nor irrelevant. They do not yield additional information about a topic but help understand the meaning of the topic. For example, in the case of positive reviews of a handbag, the words: nice, beautiful, or bag do not provide additional information, but provides better comprehension of the sentiment and topic. Hence, these are not counted towards relevancy or irrelevancy of the topic.

For the negative reviews, the same pattern emerges (Tables 6 and 7). The associated words with the first topic are almost identical in both methods. In the next topic ("Problem"), PLSA extracts more unique words that represent specific problems like "Rough", "Thread", "Material", etc., which are not present in the LSA extraction. Both models convey the information that the product does not "look"

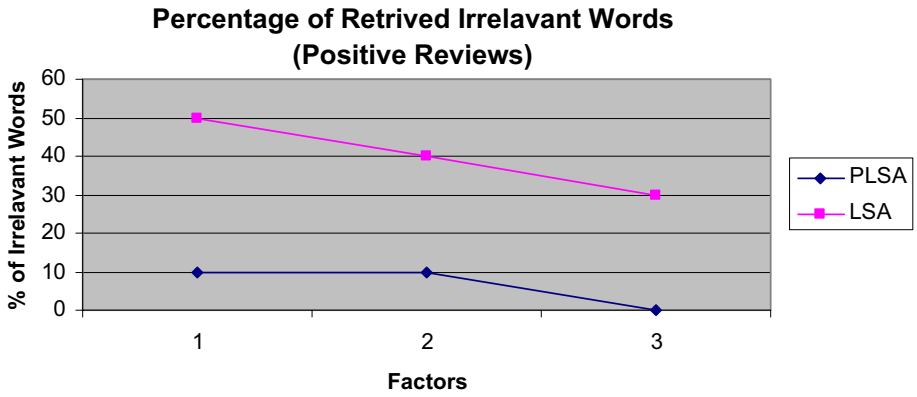

Fig. 5 Irrelevant words of positive reviews

Table 6 Negative reviews relevant words

	PLSA (retrieved relevant words)	LSA (retrieved relevant words)
Not leather	Plastic, leather, real, expect	Plastic, leather, pleather, real, expect
Problems	Material, look, photo, picture, rough, thread, leather, pleather	Photo, ugly, picture, deceive, issues
Service failure	Broken, pieces, contact, customer, help, product, quality, attach, phone, amazon	Break, pieces, cheap, faulty, phone, receive

Table 7 Negative reviews irrelevant words

	PLSA (retrieved irrelevant words)	LSA (retrieved irrelevant words)
Not leather	Spacious, pink, color	Spacious, zip, boo
Problems	Color	Seller, massive, pink, peach
Service failure		Zip, money, close

like the "picture/photo". Moreover, the service failure topic of PLSA also contains more specifics than LSA.

The precision of the two techniques for negative reviews are calculated. The Graphical representation of the precision curve is provided in Fig. 6:

The percentage of irrelevant words retrieved by the techniques is shown in Fig. 7. The graph shows that PLSA has a much lower percentage of irrelevant words than LSA (Fig. 7).

It is quite clear from these figures that LSA performs less efficiently than PLSA when analyzing reviews from a particular brand, or LSA was not able to extract the specifics to the extent that PLSA did. The real example of positive and negative reviews (Fig 8) provides supports for the superiority of PLSA in this context. LSA was not able to effectively extract the complaints in negative review and large and spacious component of positive reviews.

With this in mind, we proceed to the next analysis to see if this pattern holds in other context. We extracted topics from a broader category "Kitchen Appliances" which contains reviews of various brands and appliances. As before, we divided positive and negative reviews into two groups based on their star rating. We then extract topics from the

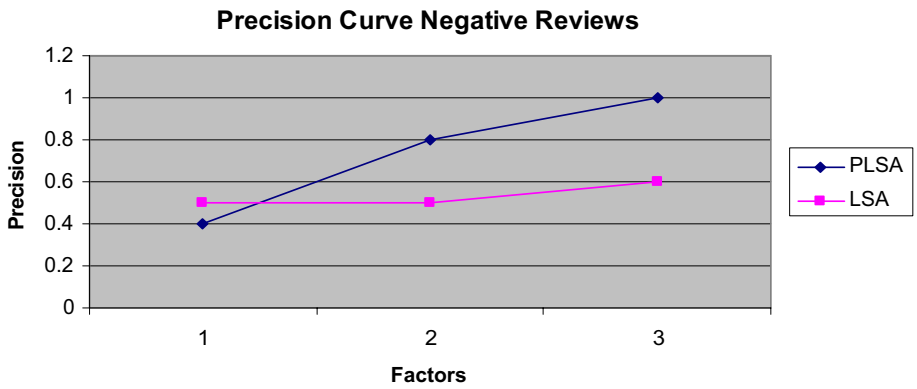

Fig. 6 Precision curve of negative reviews

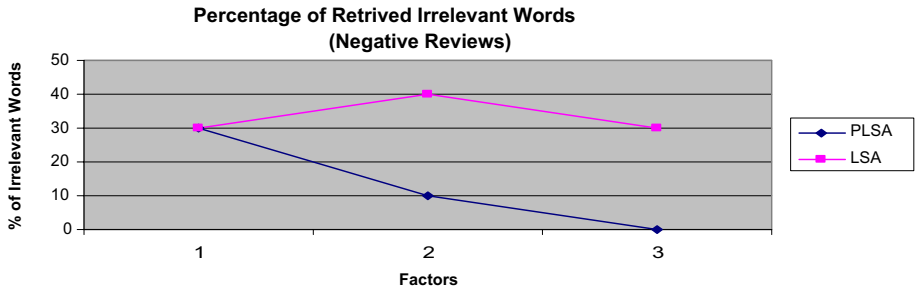

Fig. 7 Percentage of retrieved irrelevant words in negative reviews

5.0 out of 5 stars **Beyond cute!**, April 11, 2011

By
Therese A. Davis "TAD14" (ILLINOIS) - See all my reviews
(REAL NAME)

Amazon Verified Purchase(What's this?)

This review is from: Rose Handbag By Fash (Apparel)

This purse is totally cute fashion! Read other reviews that lead me to purchase it and am glad I did. Roomy - large - a real fashion statement! Color is beautiful!
Help other customers find the most helpful reviews
Was this review helpful to you? Yes No

2 of 3 people found the following review helpful:

1.0 out of 5 stars **BROKEN and UNUSABLE**, July 2, 2011

By
Jenna Nuzzi - See all my reviews

This review is from: Rose Handbag By Fash (Apparel)

Received this purse from my wish list as a gift from my Mom and the clasp to attach was broken. I contacted Amazon and they did nothing to help me and refused to contact the seller and basically told me to buy a new one. I used to be a loyal Amazon customer but they have the worst customer service ever. Years of being a customer and 1 problem and they do nothing to help me. Do not buy this purse as it is cheaply maid obviously and the company does not back their product enough to replace broken pieces. A really big waste of money I am very disappointed. How can I use a purse when the matching strap to carry the purse is broken. Very frustrating.
Help other customers find the most helpful reviews
Was this review helpful to you? Yes No

Fig. 8 Example of positive and negative Reviews of the Handbag

reviews. The results are shown in Table 8. A careful examination of the topics reveals that PLSA has formed the topics according to the specific appliances. For example, oven, pan and skillet; baking needs, then knives. On the other hand, if we look at the topics from LSA, it provides an overall summarization of the important aspects and attributes of this product category.

It can be seen that LSA extracts topics that provide information about an attribute of the product category. For example, it can be inferred by looking at the factors extracted by LSA that, customers talk about core functionalities, aesthetics, branding, technical aspects, and affective content in the reviews. However, if the topics of PLSA are examined, it is evident that the topics are extracted according to the appliances. For example, first topic relates to "oven, pan, skillet", the second one relates to "baking", the third one to "knives", and then "kettle and tea". Unlike LSA topics, these do not express core themes of the reviews. Therefore, from a managerial perspective, information in the topics extracted by PLSA has little to no use. On the other hand, the topics in LSA provide the perspective of what customers generally look for in this broader product category. For example, customers are happy if the appliances have an aesthetic attribute in addition to the core functionalities and technical superiority. Moreover, this category seems to be a popular choice for gift giving. Customers also compare different brands when buying in this product category. All this information helps a manager decide about the attributes to include in a new product in this category or improvement of the product. Therefore, in this scenario, LSA works better in terms of interpretability. The following review supports the results we received from LSA which were not visible by PLSA.

"An elegantly designed LONG WIDE toaster…..Very clean, modern appearance. Looks great sitting on the kitchen counter, whereas many of the other toaster models today look like ugly chrome spaceships from the 1950's. Personally, I'm not into that kind of retro look……." (aesthetics).

Or "This ice cream Maker is "GREAT". The fact that I can use an industrial motor (my kitchen Aid mixer) is fantastic….." (technical aspect).

"….Also makes a fabulous wedding, shower, or housewarming gift. Forget expensive wedding registries—buy the bride a lodge dutch oven and skillet. She'll hand them down to the next generation….." (gift giving/affective).

Therefore, depending on the objective of topic extraction, either PLSA or LSA becomes the superior method, and the superior performance of PLSA that was exhibited in the brand-specific reviews does not exist in every scenario. The result can be attributed to the fact that PLSA finds the highest probability terms that are likely to occur in the document. On the other hand, LSA tries to infer the topic based on the word co-occurrences.

We do not produce a performance measure curve for this section. As discussed before, the grouping of words is completely different, and a performance measure curve (or the table of relevance measurement) will not provide any

Table 8 Comparison of PLSA and LSA factors (associated words) of the positive reviews of kitchen appliances

Factors/Topics	LSA	PLSA
Core Functionalities	bake cake calphalon clean coat cook creuset dishwash easy flat fry grease handle heat muffin non nonstick oil pan set spray stainless stick stir surface ware wash wonder	absorb casserole cast cook cost creuset dish distribute dutch easily efficiency enamel flavor heat heavy iron lodge material oven pan roast season size skillet surface tend vessel whisk (oven, Pan, Skillet)
Aesthetic, more functionalities and small problem	bad bagel bake braun bread button consume counter design heat kitchen look muffin oven perfectly pick piece problem pull retro room slice slot super toast toaster whatsoever wide small	bagel bake bread buy calphalon coat cook don easy fan fry grill heat look meat muffin need non oil oven pan put scale side stick toast toaster (Baking)
Branding	analogy chef chore cutting edge cutlery differ global henckel hundred knife knives lifetime beauty blade block box brand carve nice price pro roast set sharpen sharper shear slice steak	amazon beater blade brand carve cut dough hand henckel kitchen kitchenaid knife knives mix mixer nice plastic price processor set sharpen slice spin steamer steel weight wet whatsoever whine whip (knives)
Technical aspect	fit hand kitchenaid look love mix mixer motor need potato power processor quart short speed store whip wonder attach beater big bowl cake case cloth cover cream dough	bowl box chopstick company cream customer excel fire food handle hot ice install kettle lunch machine microwave minute pour product protect remove service shear tea warranty water whistle (kettle, Tea)
Affective	awesome bar beauty calphalon clad clean color embroidery enjoy family gift haven love mattress month nonstick pillow purchase quality recommend seen set shaker sheet size skillet stainless top wedding	cocktail coffee cone cup drink enjoy fine food glass grind hours juice lose machine maker model mug press quiet read remain screw sheet spin tea thermoset top

meaningful comparison since there is no overlap of relevant and irrelevant words.

Discussion

User-generated contents are everywhere. This data contains information on sentiment and customer experiences about products or services. For market researchers, these contents are very useful and important. The use of content analysis goes back several decades in marketing. Qualitative content analysis reveals patterns, and this technique has been used in marketing for a long time (Bourassa et al. 2018; Phillips and Pohler 2019). However, contents found on the web are huge in size and usually, it is very cumbersome to manually analyze these unstructured texts. An intelligent and automated method is needed where the analysis of large amounts of data can be completed. Research has shown that competencies in big data analysis of a firm predict better performance measured by innovation, customer relationship management, etc. Big data analysis can assist in knowledge co-creation which in turn assists in making better decision (Acharya et al. 2018). More specifically, research points to the fact that domain knowledge should be incorporated while crafting the model and interpreting the result (Berger et al. 2020). Only by breaking the silos of different knowledge base, Marketing analytics can achieve its best result (Petrescu and Krishen 2021).

The current study tries to find the best method for extracting managerial information in two different marketing scenarios. Every technique has its own advantages and disadvantages. The suitability of the techniques depends on the context where it is being used. Although computer science researchers have been looking into this area for a long time, the marketing discipline started to investigate this area about a decade ago only. The knowledge and performance

measures of the techniques cannot be directly transferred to the marketing domain since the performances are context specific. For example, from a retrieval perspective (Information Technology literature), success is the system's ability to retrieve similar words or documents containing the same topic when a query word is provided to a system. So, the higher the performance, the higher the rate of finding out relevant (similar) words. On the contrary, in this marketing context, the higher the performance, the higher retrieval of the marketing manager's important information terms/documents. The current study supports the idea that the choice of a text mining approaches should be domain-specific and augmented with domain knowledge.

As mentioned before, the two contexts were different in terms of specificity, meaning that one context contained customer reviews of only one brand of handbag and the other context contained reviews of different brands and appliances of "Kitchen Products". The results show that, in the former case, PLSA extracted topics that are more meaningful and concrete. It was more interpretable and contained more information. LSA extracted topics well; but they were not as complete as PLSA topics. There were cross-words meaning that one word belonged to more than one factors. There was also a high number of irrelevant words in a topic compared to PLSA. Based on the precision and number of irrelevant words extracted by these two techniques, it can be concluded that in this context, PLSA works better in achieving the goal.

In the second context, where the goal was to learn important topics in a product category with lots of brands and products, LSA outperformed PLSA. Here also, PLSA extracted meaningful topics; but not aligned with important marketing interests. Each topic represented each appliance in the product category "kitchen appliances". More importantly, it did not group the topics according to the discussion topics of the product category (hence product attribute), which are of the main interest from a marketing manager's perspective. For example, PLSA extracted topics (Oven, Baking, Knives, etc.) may not provide a marketing manager with useful insights. It should be noted that from an information retrieval perspective PLSA might have done a fair or even superior job; however, depending on what kind of information is needed, PLSA is not a superior technique in this context. On the contrary, LSA grouped the topics according to the discussion topics of the review: core functionalities, technical aspect, branding, etc. This information is of interest to the marketing manager. Therefore, the study concludes that if the goal is to learn about a specific brand and its positive and negative attributes, PLSA reveals more specific information. However, if the goal is to learn about important aspects of a broader product category, LSA works better. The current study contributes in two ways: firstly, it responds to the recent call for research for marketing specific data analytics tool where marketing knowledge and goal is incorporated with sophisticated machine learning tools. Secondly, by experimenting in two different marketing scenarios, the study examines the suitability and superiority of two data analytics techniques.

Managerial implications

Managers can benefit greatly from understanding the topics of positive and negative reviews because they provide valuable insights into customer perceptions and preferences. By analyzing the topics that customers mention in their reviews, managers can identify areas of strength and weakness in their products, services, and overall customer experience. Using the right text mining tools, managers can identify areas for improvement. For example, the handbag should be improved in terms of its look (customers were disappointed that it did not look like leather). They can also identify areas of strength: The handbag was stylish and spacious. Managers can highlight these in their marketing messages and product descriptions, potentially driving sales and customer loyalty. Managers may also compare their product with competitors by evaluating competitors' brands. In the broader product category, the topics may reveal important aspects of the category. For example, LSA revealed that aesthetics and gift giving were important in kitchen appliances, which might not be evident. Managers can track the topics mentioned in positive and negative reviews over time and thus, can identify changes in customer perceptions and preferences.

Limitation and future research

Like any other studies, this study is not without limitation. First, for the performance measurement, the study uses a precision measure, which looks at the number of relevant words retrieved in all retrieved words. However, there are words that are relevant to the topic but not useful. For example, in the handbag positive reviews, the words "nice", "favorite" does not provide any additional information. But these words are not irrelevant words at all. To be conservative, the present study kept these words out from the "relevant" and "irrelevant" word counts so that the results are not biased. A count of irrelevant words provides another measure of performance that was used in the current study. However, the main criticism of this kind of performance measure is the subjectivity of the meaning. The precision measure is a binary approach that fails to capture the fuzziness in meaning of the words. Although the present study uses manual inspection to measure precision, the subjectivity often becomes a problem and may bias the result. To combat this problem to some extent, the ambiguous meaning words are left out while performing measurement of

irrelevant words. Another limitation was that the dataset was small. However, the size of the dataset aided the manual coding of relevant/irrelevant words that was needed to come up with precision/ recall measure. Big dataset will introduce more noise and the result may lack objectivity. As recommended in the literature, automatic text analysis can learn from manual coding of small dataset and the model can then be applied to big dataset for real-life use (Chen et al. 2018).

Application of text mining in the marketing domain is a rising phenomenon. The fact that if a text mining technique is superior in terms of information retrieval (for representing the data, retrieving similar documents, or search purposes), it might not be a superior text mining technique for a marketer's point of view. This idea warrants marketing researchers to experiment with techniques and find their suitability in different marketing contexts and needs.

Declarations

Conflict of interest The authors have no conflict of interest to disclose.

References

Acharya, A., S. Singh, V. Pereira, and P. Singh. 2018. Big data, knowledge co-creation and decision making in fashion industry. *International Journal of Information Management* 42: 90–101.

Ansari, A., and Li, Y. and Zhang, J. 2018. Probabilistic topic model for hybrid recommender systems: A stochastic variational Bayesian approach. *Marketing Science* 37 (6): 987–1008.

Berger, J., A. Humphreys, S. Ludwig, W.W. Moe, O. Netzer, and D.A. Schweidel. 2020. Uniting the tribes: Using text for marketing insight. *Journal of Marketing* 84 (1): 1–25.

Bernoff, J., and C. Li. 2008. Harnessing the power of the oh-so-social web. *MIT Sloan Management Review* 49 (3): 36–42.

Bijmolt, T.H.A., P.S.H. Leeflang, F. Block, M. Eisenbeiss, B.G.S. Hardie, A. Lemmens, and P. Saffert. 2010. Analytics for customer engagement. *Journal of Service Research* 13 (3): 341–356.

Blitzer, J., Dredze, M., and Pereira, F. 2007 Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, Prague, Czech Republic, (pp. 440–447).

Bourassa, M.A., P.H. Cunningham, L. Ashworth, and J. Handelman. 2018. Respect in buyer/seller relationships. *Canadian Journal of Administrative Sciences* 35 (2): 198–213.

Branda, A., V. Lala, and P. Gopalakrishna. 2018. The marketing analytics orientation (MAO) of firms: Identifying factors that create highly analytical marketing practices. *Journal of Marketing Analytics* 6: 84–94.

Braune, E., and L.P. Dana. 2021. Digital entrepreneurship: Some features of new social interactions. *Canadian Journal of Administrative Sciences* 39 (3): 237–243.

Chen, N., M. Drouhard, R. Kocielnik, J. Suh, and C. Aragon. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems* 8 (2): 1–20.

Chevalier, J., and D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43 (3): 345–354.

Constantinides, E., and S.J. Fountain. 2008. Web2.0: Conceptual foundations and marketinngissues. *Journal of Direct, Data and Digital Marketing Practice* 9 (3): 231–244.

Cui, D., and D. Curry. 2005. Prediction in marketing using the support vector machine. *Marketing Science* 24 (4): 595–615.

Cvitanic, T., Lee, B., Song, H. I., Fu, K., and Rosen, D. 2016. LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents. In International Conference on Case-Based Reasoning.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. Indexing by latent semantic Indexing. *Journal of the American Society for Information Science* 41 (6): 33–47.

Germann, F., G.L. Lilien, L. Fiedler, and M. Kraus. 2014. Do retailers benefit from deploying customer analytics? *Journal of Retailing* 90: 587–593.

Ghose, A., and P. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering* 23 (10): 1498–1512.

Gupta, M., and J. George. 2016. Toward the development of a big data analytics capability. *Information & Management* 53 (8): 1049–1106.

Hair, J.F., Jr., and M. Sarstedt. 2021. Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing. *Journal of Marketing Theory and Practice* 29 (1): 65–77.

Han, J., and M. Kamber. 2006. *Data mining: Concepts and techniques*. Burlington: Morgan Kaufmann Publishers.

Hanna, R., A. Rohm, and V. Crittenden. 2011. We're all connected: The power of the social media ecosystem. *Business Horizons* 54 (3): 265–273.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden.

Hou, L., Guan, L., Zhou, Y., Shen, A., Wang, W., Luo, A., ... and Zhu, J. J. 2022. Staying, switching, and multiplatforming of user-generated content activities: a 12-year panel study. *Internet Research* (ahead-of-print).

Hu, H., Zhang, C., Luo, Y., Wang, Y., Han, J., and Ding, E. 2017 Wordsup: Exploiting word annotations for character-based text detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 4940–4949).

Huang, M.H., and R.T. Rust. 2021. A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science* 49 (1): 30–50.

Iacobucci, D., M. Petrescu, A. Krishen, and M. Bendixen. 2019. The state of marketing analytics in research and practice. *Journal of Marketing Analytics* 7 (3): 152–181.

Jimenez, S., F.A. Gonzalez, A. Gelbukh, and G. Duenas. 2019. Word2set: WordNet-based word representation rivaling neural word embedding for lexical similarity and sentiment analysis. *IEEE Computational Intelligence Magazine* 14 (2): 41–53.

Kamal, A. 2015. Review mining for feature-based opinion summarization and visualization. arXiv preprint arXiv:1504.03068.

Kamps, J., and Marx, M. (2001). Words with attitude. In 1st International WordNet Conference. (pp. 332–341).

Ke, X., and Luo, H. (2015, August) Using LSA and PLSA for text quality analysis. In *2015 International Conference on Electronic Science and Automation Control* (pp. 289–291). Atlantis Press.

Kim, S., H. Park, and J. Lee. 2020. Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications* 152: 113401.

Kiron, D. 2013. Organizational alignment is key to big data success. *MIT Sloan Management Review* 54 (1): 15.

Koivisto, E., and P. Mattila. 2020. Extending the luxury experience to social media–User-Generated Content co-creation in a branded event. *Journal of Business Research* 117: 570–578.

Ku, L.-W., Liang, Y.-T., and Chen, H.-H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW)*, (pp. 100–107).

Kubler, R., Colicev, A. and Pauwels, K. 2017. Social media's impact on consumer mindset: when to use which sentiment extraction tool, *Marketing Science Institute Working Paper Series*, 17-122-09.

Landauer, T.K., P.W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes* 25 (1): 259–284.

Lee, T.Y., and E.T. Bradlow. 2011. Automated Marketing research using online customer reviews. *Journal of Marketing Research* 48 (5): 881–894.

Li, Y., and T. Li. 2013. Deriving market intelligence from microblogs. *Decision Support Systems* 55 (1): 206–217.

Li, Z., Y. Fan, W. Liu, and F. Wang. 2018. Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multimedia Tools and Applications* 77 (1): 1115–1132.

Li, S.G., Y.Q. Zhang, Z.X. Yu, and F. Liu. 2021. Economical user-generated content (UGC) marketing for online stores based on a fine-grained joint model of the consumer purchase decision process. *Electronic Commerce Research* 21: 1083–1112.

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

Liu, X., D. Lee, and K. Srinivasan. 2019. Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research* 56 (6): 918–943.

Llopis-Amorós, M.-P., I. Gil-Saura, M. Ruiz-Molina, and M. Fuentes-Blasco. 2019. Social media communications and festival brand equity: Millennials vs Centennials. *Journal of Hospitality and Tourism Management* 40: 134–144.

Lu, Y., Zhai, C., and Sundaresan, N. (2009). Rated aspect summarization of short comments. In WWW '09: *Proceedings of the 18th international conference on World Wide Web*. ACM, New York, NY, USA, (pp. 131–140).

Ma, L., and B. Sun. 2020. Machine learning and AI in marketing — Connecting computing power to human insights. *International Journal of Research in Marketing* 37 (3): 481–504.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. 2007.Topic sentiment mixture: modeling facets and opinions in weblogs. In WWW '07: *Proceedings of the 16th international conference on World Wide Web*. ACM, New York, NY, USA, (pp. 171–180).

Mikalef, P., M. Boura, G. Lekakos, and J. Krogstie. 2020a. The role of information governance in big data analytics driven innovation. *Information & Management* 57 (7): 103361.

Mikalef, P., I.O. Pappas, J. Krogstie, and P.A. Pavlou. 2020b. Big data and business analytics: A research agenda for realizing business value. *Information & Management* 57 (1): 103237.

Moe, W., and D. Schweidel. 2012. Online product opinions: Incidence, evaluation, and evolution. *Marketing Science* 31 (3): 372–386.

Moussa, M.E., M.H. Ensaf, and M.H. Haggag. 2018. A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal* 3 (1): 82–109.

Mudambi, S., and D. Schuff. 2010. What makes a helpful online review? A study of customer reviews on Amazoncom. *MIS Quarterly* 34 (1): 185–200.

Mudasir, M., R. Jan, and M. Shah. 2020. Text document summarization using word embedding. *Expert Systems with Applications* 143 (4): 111–192.

Netzer, O., R. Feldman, J. Goldenberg and Fresko, M. 2012. Mine your own business: Market-structure surveillance through text mining. *Marketing Science* 31 (3): 521–543.

Petrescu, M., and M. Krishen. 2021. Focusing on the quality and performance implications of marketing analytics. *Journal of Marketing Analytics* 9: 155–156.

Phillips, B.J., and D. Pohler. 2018. Images of union renewal: A content analysis of union print advertising. *Canadian Journal of Administrative Sciences* 35 (4): 592–604.

Ransbotham, S., C. Kane, and N. Lurie. 2012. Network characteristics and the value of collaborative user-generated content. *Marketing Science* 31 (3): 387–405.

Sahut, J.M., L.P. Dana, and M. Laroche. 2020. Digital innovations, impacts on marketing, value chain and business models: An introduction. *Canadian Journal of Administrative Sciences* 37 (1): 61–67.

Salehan, M., and D.J. Kim. 2016. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems* 81: 30–40.

Sällberg, H., S. Wang, and E. Numminen. 2022. The combinatory role of online ratings and reviews in mobile app downloads: an empirical investigation of gaming and productivity apps from their initial app store launch. *Journal of Marketing Analytics* 5: 8.

Savage, M., and R. Burrows. 2009. Some further reflections on the coming crisis of empirical sociology. *Sociology* 43 (4): 762–772.

Sidorova, A., N. Evangelopoulos, J. Valacich, and T. Ramakrishnan. 2008. Uncovering the intellectual core of the information systems discipline. *MIS Quarterly* 32 (3): 467–482.

Skeen, S.J., S.S. Jones, C.M. Cruse, and K.J. Horvath. 2022. Integrating natural language processing and interpretive thematic analyses to gain human-centered design insights on HIV mobile health: Proof-of-concept analysis. *JMIR Human Factors* 9 (3): e37350.

Steinberger, J., and Ježek, K. (2009). Update summarization based on latent semantic analysis. In I*nternational Conference on Text, Speech and Dialogue* (pp. 77–84). Springer, Berlin

Timoshenko, A., and J. Hauser. 2019. Identifying customer needs from user-generated content. *Marketing Science* 38 (1): 1–20.

Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In WWW '08; *Proceeding of the 17th international conference on World Wide Web*. ACM, New York, NY, USA, (pp. 111–120)

Turney, P., and M.L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transaction Information Systematic* 21 (4): 315–346.

Vollrath, M., and S. Villegas. 2022. Avoiding digital marketing analytics myopia: Revisiting the customer decision journey as a strategic marketing framework. *Journal of Marketing Analytics* 10: 106–113.

Vorvoreanu, M., G. Boisvenue, C.J. Wojtalewicz, and E. Dietz. 2013. Social media marketing analytics: A case study of the public's perception of Indianapolis as Super Bowl XLVI host city. *Journal of Direct, Data and Digital Marketing Practice* 14 (4): 321–328.

Wamba, F., A. Gunasekaran, S. Akter, S.. Ji.-fan Ren, R. Dubey, and S.J. Childe. 2017. Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research* 70: 356–365.

Yang, M., Y. Ren, and G. Adomavicius. 2019. Understanding user-generated content and customer engagement on Facebook business pages. *Information Systems Research* 30 (3): 839–855.

Yu, X., Y. Liu, X. Huang, and A. An. 2012. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering* 4 (4): 720–734.

Zhang, K., T. Evgeniou, V. Padmanabhan, and E. Richard. 2012. Content contributor management and network effects in a UGC environment. *Marketing Science* 31 (3): 433–447.

Zhong, Ning, and David A. Schweidel. 2020. Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science* 39 (4): 669–686.

Zhu, L., Gao, S., Pan, S. J., Li, H., Deng, D., and Shahabi, C. (2013) Graph-based informative-sentence selection for opinion summarization. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 408–412).

Shimi Naurin Ahmad is an associate professor in the Department of Business Administration at Morgan State University. She holds an M.S. in Electrical Engineering and Ph.D in Marketing from Concordia University, Canada. Her research interests include online consumer behavior and application of data/text mining techniques in marketing. She has presented her work at conferences such as the Academy of Marketing Science and Advanced Research Techniques Forum. Her work has appeared in the *Journal of Business Research, International Journal of Information Management, International Journal of Electronic Commerce, Journal of Marketing Analytics* among others.

Michel Laroche holds a Ph.D. and M.Ph. (Columbia), D.Sc.*hc* (Guelph), and M.Sc.Eng. (Johns Hopkins). He is a Fellow of Royal Society of Canada, APA, SMA, and AMS. He received the 2016 *Hans B. Thorelli Award, 2019 Howard Berkman Service Award*. He published 200 articles in the *Journal of Consumer Research, Journal of Business Research, Journal of the Academy of Marketing Science, Journal of the Association for Consumer Research,* and *International Journal of Information Management*. His projects involve consumer behavior, digital marketing, brand communities, social media, and sharing economy. He is the Editor-in-chief of the *Canadian Journal of Administrative Sciences.*