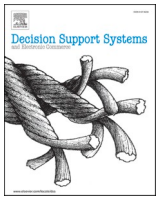
Decision Support Systems 157 (2022) 113751

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Sourcing product innovation intelligence from online reviews

David M. Goldberg [a,*], Alan S. Abrahams [b]

[a] Department of Management Information Systems, Fowler College of Business, San Diego State University, San Diego, CA 92182, United States of America
[b] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, United States of America

ARTICLE INFO

Keywords:
Online reviews
Text mining
Data mining
Innovation
Business intelligence

ABSTRACT

In recent years, online reviews have offered a rich new medium for consumers to express their opinions and feedback. Product designers frequently aim to consider consumer preferences in their work, but many firms are unsure of how best to harness this online feedback given that textual data is both unstructured and voluminous. In this study, we use text mining tools to propose a method for rapid prioritization of online reviews, differentiating the reviews pertaining to innovation opportunities that are most useful for firms. We draw from the innovation and entrepreneurship literature and provide an empirical basis for the widely accepted attribute mapping framework, which delineates between desirable product attributes that firms may want to capitalize upon and undesirable attributes that they may need to remedy. Based on a large sample of reviews in the countertop appliances industry, we demonstrate the performance of our technique, which offers statistically significant improvements relative to existing methods. We validate the usefulness of our technique by asking senior managers at a large manufacturing firm to rate a selection of online reviews, and we show that the selected attribute types are more useful than alternative reviews. Our results offer insight in how firms may use online reviews to harness vital consumer feedback.

1. Introduction

This paper explores how firms can capitalize upon feedback from online reviews to derive vital insights that assist with product innovation. Manufacturers are constantly looking for ways to improve upon existing product lines or to branch into related products that can improve their competitive position. Efforts to innovate existing product lines are often nonlinear in the sense that new ideas may suddenly spur on unforeseen changes and improvements [49]. However, determining the most effective new ideas on which to build and ensuring that these ideas align with consumer demand proves difficult and complex. Ideas for product innovation may come from many sources, including brainstorming, competitive monitoring, focus groups, warranty claims, and online media [44]. In this paper, we consider innovation from a knowledge management perspective. As defined by du Plessis [14], we consider innovation to be "*the creation of new knowledge and ideas to facilitate new business outcomes, aimed at improving internal business processes and structures and to create market driven products and services.*" In this study, we consider approaches to develop firms' understanding of their own products' key attributes, both positive and negative, relative to those of their competitors.

MacMillan and McGrath [34,35] propose the widely-accepted attribute mapping framework for interpreting and prioritizing innovation opportunities. The framework distinguishes between positive, negative, and neutral attributes based on simple consumer preference and then between basic, discriminator, and energizer attributes based upon consumers' likelihood to make purchasing decisions due to that sentiment. Although this framework has provided managers with a useful means of differentiating potential product attributes for decades, research has generally taken the form of case studies rather than large-scale statistical analyses [2,38,52].

In the years since the advent of the attribute mapping framework, firms have been faced with a new challenge in taking advantage of online media platforms. Online word-of-mouth has expanded enormously, and thousands of posts on social media and in the form of online reviews each day offer new feedback on consumers' product preferences. There is evidence that more positive online reviews are associated with greater sales [10], as 91% of consumers read online reviews to better understand products and make purchasing decisions [7]. Amidst this avalanche of feedback, researchers and practitioners alike have sought to understand how to extract the most useful information [29]. Unfortunately, the incredible volume of online feedback makes manual review

* Corresponding author.
E-mail address: dgoldberg@sdsu.edu (D.M. Goldberg).

https://doi.org/10.1016/j.dss.2022.113751
Received 2 April 2021; Received in revised form 16 December 2021; Accepted 4 February 2022
Available online 9 February 2022

of each post unreasonable [1,20]; however, harnessing the incredible volume of online feedback to make the most critical of the requested product innovations could offer firms fantastic competitive advantages.

In this paper, we use the attribute mapping framework to drive rapid and automated prioritization of online reviews. We make use of a large dataset of manually coded Amazon.com reviews [37] and perform a series of text mining methods to differentiate those reviews containing vital innovation feedback. We adapt the heuristic text mining method proposed in Goldberg and Abrahams [19] for the detection of safety hazards in online reviews, and in this study, we apply the technique to detection of innovation opportunities. We use this method to develop lists of terms that identify reviews of interest to innovation, such as complaints or complements about key product attributes and requests for new product features. We aim to provide a curated shortlist of reviews that provide value for firms to consider as they evaluate opportunities to improve product offerings. We compare our techniques to existing state-of-the-art text mining methods, such as sentiment analysis [25,41], aspect mining [3,45], and Latent Dirichlet Allocation (LDA) [5], and we find that our method outperforms these techniques. The usefulness of the insights in the online reviews retrieved are verified by an assessment from senior-level managers at a large Fortune 1000-listed manufacturer of countertop appliances earning over $500 million in revenue per year.

This paper makes key contributions to several different research streams. First, this paper contributes to the theoretical stream of research on the attribute mapping framework [34,35] by providing empirical support for the theory's application to practice. As a final stage of analysis, we asked senior-level managers at the collaborating countertop appliance manufacturer to rate the usefulness of the reviews identified by our technique versus a random chance baseline. Each category of reviews was rated as significantly more useful than alternative online reviews. Our results show not only that the attribute mapping framework can be applied to consistently select different categories of online reviews, but also that those online reviews provide meaningful insights that assist product designers and managers in their innovation efforts. Second, this paper provides the first method for harnessing the vast volume of online reviews to provide rapid business intelligence targeted at innovation opportunities. Prior research has examined the extraction of basic product attributes from online reviews [29], but this paper is the first to bridge the gap between the discussion of attributes in these online discussions and categorizing the specific feedback for product designers to respond to the most pressing innovation opportunities (compliment, feature request, or irritator). The immense volume of online reviews presents a difficult challenge for modern firms to navigate [1]; automated tools that cut through the intimidating volume of online content and prioritize the important insights save time and focus innovation on the most important areas. Third, our study provides actionable insights for the countertop appliances industry. We present a series of words and phrases that distinguish reviews of interest to this industry that firms can implement with immediate effect to prioritize content.

The remainder of this paper is structured as follows. In our literature review, we explain the theoretical foundations for our study from MacMillan and McGrath [34,35], and we contextualize their studies in the innovation and entrepreneurship literature. We also discuss recent literature on online reviews and media as well as the challenges associated with extracting information from these formats. We describe the key research questions that we seek to answer in this work as well as its contributions. Then, we detail our methodology, including the dataset employed in this work, the algorithms used to prioritize the most important online reviews, and the competing techniques to which we will compare our findings. We detail the results generated from our technique and competing techniques. We validate the usefulness of our findings by a comparison performed by senior-level managers at the collaborating countertop appliances manufacturer. Finally, we conclude our paper, noting its implications as well as the potential for future work.

2. Literature review

A key tenant in entrepreneurship is improving offerings within the firm's product line to differentiate it from competitors and to offer a competitive advantage. Much of the entrepreneurship literature confirms this notion by arguing for an association between innovation and the success or profitability of a firm's entrepreneurial ventures [33,47]. There is also empirical evidence that small and medium firms can improve their profitability by adapting to changes in their business environment and innovating faster than their competition, for whom these reactions may be more difficult [47,56]. Innovation pressures have increased in recent years, as product and business model life cycles have shortened, increasing the need for firms to adapt quickly to stimuli in order to stay competitive [42]. Firms that do not source inspiration for new ideas quickly and capitalize upon profitable solutions may quickly be left behind. The literature acknowledges a distinction between two types of firms: firms that generate new solutions and technologies internally and firms that adopt or build upon those solutions or technologies advanced by other firms [11]. The literature uses several different monikers to characterize this distinction, such as innovation-generating versus innovation-adopting [11], innovators versus imitators [9], or first movers versus second movers [42]. However, Pérez-Luño, Wiklund and Cabrera [42] acknowledge that these different types of firms present more of a continuum than a dichotomy; in practice, most firms initiate some original ideas internally while also monitoring their competitors to ensure that their products or services do not lack attributes standard in the space.

The literature makes a key distinction between novelty and innovativeness [30]. Jackson and Messick [24] describe "novelty" as the extent to which a concept or instantiation differs from convention. Building upon that definition of novelty, Sethi, Smith and Park [50] describe "innovativeness" as encapsulating whether an idea is "different from competing alternatives in a way that is valued by customers." A key distinction between these definitions is that usefulness is not a necessary component of novelty; an idea can be novel in the sense that it differs from convention without being viewed as desirable by the consumers whose needs it is ultimately meant to satisfy in application. Therefore, differentiation of potential novel ideas is vital to successful firms. Ideas that are novel without being innovative may consume resources to a greater extent than they provide revenues [17], but innovative ideas can have a transformative impact that propels firms ahead of the competition [32]. Therefore, it is of vital importance for firms to organize and prioritize their ideas so that they can focus on the strategies that provide the best fit with consumer preference.

Several prior works have examined the potential for using online data to improve understanding of consumer preferences; however, there are still substantial gaps in the literature. For example, Qi, Zhang, Jeon and Zhou [45] and Amplayo, Lee and Song [3] each implement aspect-based models in effort to mine consumer requirements. The modeling includes an initial topic modeling layer followed by sentiment analysis to determine the polarity of consumer opinion on each topic. A significant advantage of this unsupervised approach is that it is simple and inexpensive to implement. However, a disadvantage is that the unsupervised nature of the approach complicates whether the topics identified truly differentiate each product relative to the competition. While the methodology can determine which topics are mentioned in reviews and whether they are mentioned positively or negatively, it does not contain the specificity to identify which aspects of products are most associated with purchasing decisions, a question that we hope to answer in this work. For instance, consumers may feel negatively about certain aspects of products, but if they perceive that the product's competitors also perform poorly on those aspects, then purchasing decisions may not be affected. In addition, while the methodology assesses associations with the helpfulness of online reviews in the eyes of other consumers,

this work will go beyond this to assess helpfulness in the eyes of industry innovators, who can evaluate the real-world feasibility of consumer requests. Siering, Deokar and Janze [51] present a similar work in identifying service requests in airline reviews. This paper utilized word prevalence to determine an initial set of service-related topics, then utilized sentiment to determine polarity associated with each topic. For this work as well, the lack of a supervised model suggests that it is possible to tell which features consumers like or dislike but not the extent to which each feature drives purchasing decisions. Xu, Liao, Li and Song [59] evaluate comparative statements between products, such as "N95 has better reception than Motorola RAZR2 V8 and Blackberry Bold 9000." Though these statements are valuable feedback, they are very specific, and thus the technique would omit a great deal of other valuable feedback. Zhang, Fan, Zhang, Wang and Fan [63] focus not on existing product features but rather on potential improvements proposed by customers. Unlike the prior papers mentioned, the authors utilized a supervised approach to classify reviews on the sentence-level, seeking to identify sentences mentioning these potential improvements using deep learning models. However, one limitation of this work is that it focuses only on potential new features and not on the relative merits of existing features. In addition, the usefulness of these derived features is in question as some consumer suggestions may not be feasible to implement given industry constraints.

2.1. Theoretical underpinnings

MacMillan and McGrath [34,35] provide a model known as the attribute mapping framework for helping firms prioritize and differentiate product attributes. The attribute mapping framework has been widely applied to a variety of domains, generally in the form of case studies, such as applications in the pharmaceutical industry [38], e-business [2], and family firms [52]. The framework delineates between two dimensions of product attributes. On the horizontal axis, the framework lists positive, negative, and neutral attributes, where positive attributes are desirable to consumers, negative attributes are undesirable to consumers, and neutral attributes do not affect consumer purchasing decisions. Importantly, however, not all attributes within a row are equally influential. The vertical axis delineates between basic attributes, discriminator attributes, and energizer attributes. Basic attributes are important to consumers but unlikely to be a source of product innovation; regardless of the sentiment that consumers have concerning a specific attribute, "basic" status implies that these attributes reflect fundamental expectations of consumers. For example, consumers may view a computer's ability to play videos as a positive attribute; however, this feature is now so ubiquitous that virtually all consumers expect video playback as a standard feature of any new computer. Discriminator attributes may be a source of innovation; they imply that some specific group of consumers may view the given attribute as reason to choose one product over a competitive product, although this differentiation may not apply to all consumers. Product color is an example of a negative differentiator (dissatisfier); some consumers may choose a competing clothing brand if they cannot find a shirt in their favorite color. However, not all consumers will necessarily share that preference, so the dissatisfaction is only relevant among some subset of consumers. Finally, energizers are vital attributes that have near-universal reactions from consumers and create great vigor and motivation within a consumer base to make a certain purchasing decision. Revolutionary technologies may frequently fall into the category of positive energizers (exciters); for example, Apple's original iPhone, which unified the mobile phone, media consumption, and touch display technology in a single device, offered consumers such an exciting new alternative to contemporary mobile phones that it invigorated and inspired many users to purchase iPhones. Table 1 displays the attribute mapping framework in tabular format, adapted from MacMillan and McGrath [34,35].

The attribute mapping framework has several useful applications for managing business innovation. The first such application is building a basic profile of a product's strengths and weaknesses with a particular understanding of the attributes of the product that are most important to consumers. Product designers and managers generally build an intuitive understanding of the strengths and weaknesses of their own products, but it is often difficult to separate their personal feelings and evaluate these products without substantial bias, necessitating the need for consumer-driven innovation [4,13]. Often, sourcing information on product attributes from outside a product development team reveals that consumers evaluate those products differently than the internal team, and understanding these preferences better allows firms to be responsive to demand [39]. Relatedly, this step allows firms to verify their assumptions about a product's desirability. For example, comparing the attribute maps of multiple products in the same category can allow firms the ability to understand why a consumer might choose one product over another. Clarifying the reasoning for purchasing decisions maximizes a firm's ability to adapt and respond to these preferences. Perhaps the attribute mapping framework's most valuable feature is that it provides a means of differentiating between and prioritizing different innovation opportunities. Neutral product attributes are not a promising area for major investments in product development in that consumers do not feel strongly about them in a positive or a negative sense. Altering or improving neutral attributes does not improve a product's position substantially relative to the competition because purchasing decisions are unaffected. Positive and negative attributes may or may not be important areas of emphasis; basic positive or negative attributes may be appreciated or dissatisfying to consumers respectively, but as they represent standard expectations for the product, they are not a substantial way in which to sway purchasing decisions. Instead, the most vital parts of the attribute map for prioritizing product development efforts are positive or negative discriminators or energizers, as these attributes result in different purchasing decisions. MacMillan and McGrath [34,35] suggest that firms ought to emphasize positive energizers (exciters) and rectify negative energizers (enragers) first, as these attributes almost universally affect purchasing decisions. Positive and negative discriminators also require attention, particularly when they pertain to a large subset of consumers.

Table 1
Attribute mapping framework. Adapted from MacMillan and McGrath [34,35].

	Basic	Discriminators	Energizers
Positive	Non-negotiables *Performs at least as well as competition*	Differentiators *Performs better than competition if attribute is salient to target customers*	Exciters *Performs better than competition*
Negative	Tolerables *Performs no worse than competition*	Dissatisfiers *Performs worse than competition if the attribute is salient to target customers*	Enragers *Performs worse than competition; must be corrected at any cost*
Neutral	So whats? *Does not affect purchasing decision in a meaningful way*	Parallel differentiators *Influences segment attitudes but is not directly related to performance*	N/A

2.2. Online reviews and media

Given the proliferation of the Internet in recent years as a vibrant form of interpersonal communication, consumers look online more and more for an understanding of the products that they may be interested in purchasing. The exchange of information by users concerning their sentiments and experiences with products is referred to as word-of-mouth (WOM) [63]. The literature suggests not only that consumers read WOM frequently as a means of gathering information about a product of interest, but also that they frequently make purchasing decisions based upon WOM [7,10]. There are many potential venues for WOM, including social media sites (Facebook, Twitter, etc.), online review sites (Amazon, Target, Walmart, etc.), various online forums, and

more. Types of WOM may vary by source; for example, Facebook posts are rather free-formed and may or may not present specific product feedback, whereas online reviews are more targeted to specific products and manufacturers. The growth of online reviews has been explosive in recent years, as the world's largest online review platforms now contain hundreds of millions of reviews [37].

Given the enormous volume of product feedback conveyed through online reviews, they present a compelling new source of information for firms. Product designers often consider many data sources when revising their products, including focus groups, consumer complaints and warranty claims, and their own ingenuity [44]. While not all consumer suggestions are feasible, experimental evidence suggests that product design that considers consumer feedback out-sells product design performed in laboratory [39]. In fact, various academic outlets have called for methods that emphasize consumer-driven innovation [4,13] as there is empirical evidence that responding to consumer preferences improves satisfaction and thus has the capacity to improve competitive positioning [62]. While many practitioners surely make use of online reviews in their product development processes, online reviews are so voluminous that it is nearly impossible for practitioners to systematically read them all. Therefore, methods of analyzing the content of online reviews in an automated fashion offer the possibility of enormous time savings for many firms [1,19], and they ensure that firms are able to actually process and respond to the most critical online feedback.

Text mining has become a popular emerging field for researchers and practitioners alike to extract meaning from online data sources. Analyzing textual data poses some unique difficulties in the sense that it does not conform to the differentiable fields of tabular-formatted data; instead, each record is lumped into a text field. Further, many aspects of the English language (and other languages) are idiosyncratic, representing a further challenge for algorithmic approaches. Many sentiment approaches operate based on a "dictionary" of terms trained with a machine learning model that distinguishes positive text from negative text [25,41]. These dictionaries are used to rate unseen text on a sentiment scale, where positive values such as +5 typically denote positive emotive content and negative values such as −5 typically denote negative emotive content. Applications of sentiment techniques have also been broad, such as distinguishing between positive and negative consumer attitudes [18,54] or predicting the stock market's response to sentiment on social media platforms [6]. Text mining methods have generally been effective at distinguishing between positive and negative consumer sentiment in online media [18,54].

Despite its obvious appeal for analyzing online reviews and other types of online media, researchers have been quick to note some of the clear limitations of sentiment analysis [19,28]. First, because sentiment techniques tend to rely on pre-built dictionaries, they are often ineffective at coping with some of the linguistic exceptions to common conventions in the English language. For example, sentiment dictionaries typically label the word "awful" as invoking strongly negative sentiment. However, online text may use phrases such as "the price was awful good," which instead actually invokes strongly positive emotive content. Second, sentiment techniques are rather blunt tools in that positive or negative sentiment does not necessarily imply rich or interpretable content. For example, the sentence "the product was terrible" would correctly be identified by most sentiment techniques as expressing negative sentiment. However, the sentence is nonspecific and does not offer any meaningful feedback that the firm could use for remediation efforts. This problem may be compounded by domain-specific language, such as an appliance leaking. Non-emotive or neutral terms like "leaking" offer the necessary specificity to be actionable for firms, as opposed to more emotive terms like "terrible" that do not specify a problem. Therefore, despite the appeal and success of sentiment analyses for some applications, they possess some notable limitations and deficiencies due to which more nuanced techniques may offer superior performance.

2.3. Framework for innovation opportunity discovery

We make several adaptations to the attribute mapping framework for use in the study of online reviews. The first and most fundamental change is to combine some of the initial attribute types. Using text mining techniques, we hope to identify key words and phrases that distinguish reviews that are discussing attributes of interest to firms' product innovation. However, the literature acknowledges that product feedback in the form of online reviews is not without bias; in particular, self-selection bias affects some of the content in online reviews [22,31]. That is, as opposed to feedback being solicited from a random and representative sample of consumers, an online review platform is a forum in which the consumers elect to participate outside of the firm's direct control. This facet of online reviews has both advantages and disadvantages. One advantage of this type of feedback is that consumers that self-select are likely to be highly motivated to share their opinion about a product [22] and as such will tend to offer more detailed and motivated product feedback than the average consumer. Whichever attributes of a product an online reviewer discusses in their review are likely attributes of great importance in their view. A necessary consequence is that this context makes distinguishing between discriminators and energizers quite difficult. Recall that discriminators may cause consumers to choose one product over another only if the relevant attribute is salient to them; in contrast, energizers represent more universal differentiation and motivation. As online reviewers self-select, they are inherently consumers to whom the attributes they discuss are salient, and thus within a review it is difficult to differentiate between a discriminator that only applies to a subset of consumers and an energizer that applies more broadly. One method for delineating discriminators from energizers post-hoc may involve counting the instances of a given attribute's mention; attributes that are mentioned most often are more likely to have broad interest (energizers), while attributes that are mentioned less often may only have narrower interest (discriminators). However, we note that this is not always the case, as the frequencies of mentions in reviews pertain not only to the value that customers place certain features, but also on how likely they are to have had a certain experience. For example, while a product safety hazard may be an energizer in the sense that it would evoke strongly negative reactions for nearly all consumers, many safety hazards due to manufacturing defects may be relatively rare and only affect a few consumers [61]. As such, frequency of mentions in online reviews alone would not be sufficient to delineate this energizer from a discriminator. Thus, for the purposes of the analysis of a single review, we combine discriminators with energizers, as a single consumer's feedback does not provide enough information to distinguish between the two. We refer to positive discriminators/energizers as "compliments" and to negative discriminators/energizers as "irritators."

A further adaptation upon the attribute mapping framework for use in analyzing online reviews is the addition of feature requests. The literature on mobile apps acknowledges that mobile app developers respond to several different types of feedback in online reviews, namely bug reports in which the developers are asked to fix a specific problem and feature requests in which the developers are asked to add a specific function [23,46]. Feature requests represent a unique extension of the attribute mapping framework. Applied to the wider ecosystem of products, feature requests represent instances in which a consumer proposes a new feature or addition be added to a product in a future iteration. Feature requests differ from compliments and irritators, both of which refer to preexisting attributes of products, because feature requests refer to potential attributes of products that are not yet implemented. Feature requests may be either positive or negative: a consumer may request a new feature that rectifies an existing problem with the product or a feature that adds a desirable new dimension to the product.

MacMillan and McGrath [34,35] note the use of the attribute mapping framework to evaluate new ideas for potential attributes, but methods for sourcing these innovation ideas are often expensive,

depending upon focus groups or consumer surveys. Exploring feature requests in online media offers product development teams actionable information based on the real-time opinions of thousands of consumers. In Table 2, we display our adapted attribute mapping framework, which is updated to reflect compliments, feature requests, and irritators. As these new types of attributes are the most striking for firms to rectify, improve upon, or otherwise consider, we focus on these core types of attributes in our text mining study.

3. Research questions and contribution

In this paper, we seek to address three key research questions. First, how prevalent are compliment, feature request, and irritator attributes in online reviews? Second, to what extent can text mining techniques be employed to extract or prioritize these attributes in online reviews, and which text mining techniques perform best? Third, to what extent are the insights derived from these attributes useful to product innovators in practice?

We make three key contributions in this study. First, this study proposes an empirical validation of MacMillan and McGrath [34,35]'s attribute mapping framework, which offers a theoretical view of product innovation opportunities, but evidence of the framework's effectiveness has been anecdotal or in the form of case studies rather than a large-scale empirical validation [2,38,52]. We examine online reviews and characterize their postings with respect to the prevalence of these important attribute types. Our findings suggest the viability of online reviews as a potential source for innovation ideas to supplement existing approaches, such as internal brainstorming or focus groups [44]. The proposed study is a novel use of state-of-the-art text mining methodology in this seldom-explored arena and a vital empirical validation of a long-standing theory in the literature. Second, this study is the first to leverage online reviews for automated extraction of innovation opportunities. Perhaps some of the most similar work was performed by Lee and Bradlow [29], who develop a text mining model for extracting marketing data from online reviews. The authors' model focuses on which product attributes are being discussed in online reviews, but it does not extend to the level of identifying the type of feedback (e.g., feature request or compliment). Additionally, the existing model is aggregated, focusing on general trends across many reviews rather than determining which individual reviews may be most interesting. The authors exhort researchers to specifically pursue data mining on "user needs" in online reviews, a call which has not yet been filled. This study will emphasize prioritization of user needs: which reviews are the most pertinent to the feedback attribute type of interest (irritators, feature requests, or compliments). Using these techniques, practitioners can narrow the process of soliciting feedback from online reviews down to a smaller sub-sample of only the most interesting information. Third and finally, this paper presents actionable insights with immediate industry application for the countertop appliances industry. The techniques discussed in the paper and the distinctive terms generated to detect product attributes can be applied with immediate effect in this industry for the purpose of monitoring online reviews.

Table 2
Revised attribute mapping framework for interpreting online reviews. Adapted from MacMillan and McGrath [34,35].

	Basic	Discriminators	Energizers
Positive	Non-negotiables *Performs at least as well as competition*	Compliments *Performs better than competition and is salient to target customers*	
Negative	Tolerables *Performs no worse than competition*	Feature requests *Potential to improve performance relative to competition* Irritators *Performs worse than competition and is salient to target customers; must be corrected at any cost*	
Neutral	So whats? *Does not affect purchasing decision in a meaningful way*	Parallel differentiators *Influences segment attitudes but is not directly related to performance*	N/A

4. Methodology

In the following, we describe the methodological steps utilized in this paper. Broadly, we began with a large dataset of Amazon.com reviews, which we began by "tagging" or labeling manually for the three target classes: irritators, compliments, and feature requests. This tagging was performed by many individuals, but a randomly overlapping section and expert-validated section ensure the quality of the labeling. Thereafter, we implemented a heuristic-based text mining methodology to identify words and phrases most associated with each of the three target classes. We compare this to competing techniques in topic modeling and sentiment analysis.

4.1. Dataset and data coding

We chose Amazon.com, the world's largest e-commerce platform and largest online review platform, as the data source for this paper [37]. In collaboration with a large Fortune 1000-listed manufacturer of countertop appliances earning over $500 million in annual revenue, we chose the Amazon product categories pertaining to that firm's key product offerings and collected 733,411 total reviews pertaining to those categories. In addition to each review's basic textual content, our dataset also included the product that each review referred to, its date, its title, and its star rating on a scale of 1 to 5, inclusive. For the first stage of tagging (coding), we sought to label whether each of the reviews referred to each of the attribute types on the adapted attributed mapping framework discussed previously. We randomly chose 25,000 reviews out of this large subset for analysis. In the following, we created a scheme for delineating between online reviews that referred to feature requests, irritators, or compliments:

1) **Feature request**: The consumer is explicitly asking that the manufacturer add a specific new feature to the product or a specific modification of the product to improve the product.

Example: "I like this blender but I still need something to hold it and secure it on the base. Sometimes I feel that the top will fall off and I will be in a big trouble…"

2) **Irritator**: The consumer is unhappy or dissatisfied with a specific aspect of the product. Irritators are anything specific that the consumer dislikes or specific things that make the consumer irritated, dissatisfied, enraged, terrified, or disgusted.

Example: "ONE BIG FLAW: handle is poorly designed and has sharp ridges, so its very uncomfortable, especially when full of liquids and heavy."

3) **Compliment**: Consumer expresses happiness or satisfaction about specific aspects of the product. Compliments are specific aspects of the product that the consumer is joyful or excited about or that differentiate the product from competitor products.

Example: "I am so glad that this product came with such a nice motor… Smooth blending ability at a fraction of the cost of the more trendy blenders."

The 25,000 randomly selected reviews were randomly distributed and presented to undergraduate business students at a major public research university for tagging. Each student was asked to tag a maximum of 200 reviews for this phase of the project to avoid a loss in data quality due to tiredness or overworking. As each review was

presented to the students in an online interface, students were asked to indicate next to each review whether it referred to a given attribute type of study. To avoid the potential cognitive overload of forcing taggers to search for multiple attribute types at the same time, we staggered the tagging for each attribute type of study. That is, taggers were initially asked to identify only whether a review referred to a feature request; later, taggers were asked to identify only whether a review referred to an irritator. The separation of the tagging into binary attributes simplifies the tagging task and ideally improves the reliability of the tagging product. As the same 25,000 reviews were tagged multiple times, it is possible for one review to be labeled as referring to multiple attribute types. For example, a review could reflect both an irritator and a feature request: "the blender pitcher was awful for real cooking because it gets stuck and has to be reset. They should make a mode that periodically unsticks the blades."

An important characteristic of the above tagging scheme is that each feature request, irritator, or compliment must be specific and explicit. Many consumer reviews may reveal some generic sentiment that a consumer has regarding a certain product, but nonspecific sentiment is not an actionable component of the attribute mapping framework. For example, some reviews might nonspecifically state that "this blender was awful, definitely don't buy it." Although the reviewer clearly complains about the quality of the product, as they do not express a specific issue with the product beyond their general sentiment, the review is of little use to manufacturers. Taggers were instead encouraged to focus only on specific feedback from consumers, which is more useful for innovation purposes.

The data tagging process for this project was examined on several levels to ensure the reliability of the output. First, as reviews were assigned to taggers at random, some reviews are tagged multiple times. In these instances, it is necessary to reconcile multiple tags for one review, and particularly in the case of conflicting tags, to render a final decision. Following the example of prior work [19,57], we follow a "majority conservative" decision rule, in which we choose the majority vote in the event of any disagreement. However, if the votes are tied, then we choose the target classification (i.e., compliment, feature request, or irritator). This strategy avoids false negatives when there is any doubt as the correct code for any review. The overlapping of taggers provides us an opportunity to assess reliability because we can compute the percentage of overlapping tags on which taggers agreed with one another. A higher percentage of the overlapping tags on which the taggers were unanimous implies more reliable tagging. Second, for external validation, we asked for tagging support from senior-level managers at the collaborating countertop appliances manufacturer. These expert taggers serve as authorities to validate the tagging of our students. We asked six senior-level managers to provide tags on each of the three projects (compliments, feature requests, and irritators), assigning two managers to each project, and we provided each with the same tagging protocol as the students. Each authority tagger was supplied with a subset of reviews that students had tagged; the subset was selected using a stratified random sample such that each binary classification was equally likely. In addition to each binary classification, we also asked the managers to rate how useful they felt the review would be to their innovation process on a three-point scale [50]: (1) not useful at all; (2) a bit useful; and (3) very useful. In doing so, we have a mechanism by which to assess the real-world applicability of these reviews to innovation processes. We can compute two statistics to measure the agreement between student taggers and authority taggers: first, the percentage of tags on which the student taggers and the authority taggers agreed, and second, Cohen's K, which compares the agreement percentage to random chance agreement. A higher agreement percentage and, relatedly, a Cohen's K value closer to 1 indicate more reliable tagging. In Table 3, we present these statistics across the three tagging projects. Each project scored high levels of agreement on all measures, suggesting that the taggers consistently identified the same types of reviews in each attribute.

4.2. Term curation

In the text mining literature, Abrahams et al. [1] and, later, Goldberg and Abrahams [19] describe a supervised learning approach for curating "smoke terms" that delineate reviews that mention product safety hazards from reviews that do not. "Smoke terms" refer to distinctive words and phrases that occur especially prevalently in reviews that contain these concerns. This technique has been applied with great success for detecting product safety hazards across several industries, including baby cribs [40], dishwashers [28], and healthcare [60]. As the lexical properties of online reviews in each industry tend to differ, smoke term lists are typically tailored to a specific domain. For example, the term "choking hazard" may be a distinctive issue in baby cribs industry [40], but it does not likely generalize to dishwasher reviews.

A substantial approach of smoke terms is their interpretability, as specific words and phrases can be identified that illustrate the rationale behind a recommendation. While many sophisticated machine learning architectures offer great accuracy, this can be at the expense of users understanding the reasoning behind recommendations [26,43]. Recent research has found that such approaches can therefore threaten adoption [48]. Smoke terms retain great interpretability, improving the likelihood of practitioner acceptance. In this paper, we adapt the aforementioned smoke term methodology for use in detection of product innovation opportunities in online reviews. A schematic of this methodology is displayed in Fig. 1. We use the technique multiple times to generate lists of applicable terms for each attribute type of interest (compliments, feature requests, and irritators). We initially separate our dataset into three approximately equally sized portions. The first portion (A) is the training set, which is used to generate a basic linguistic understanding of the online reviews and determine candidate terms to distinguish features of interest [15,19]. The second portion (B) is the curation set, which is used to test the candidate smoke terms and choose those that provide the highest level of performance. The third portion (C) is the validation set, which is used to provide measures of the efficacy of the term lists.

After our dataset is divided into three portions, we use information retrieval techniques to identify candidate smoke terms that may distinguish reviews of interest to a given attribute type. Each review is classified into two binary categories; for example, "feature request" versus "no feature request" or "irritator" versus "no irritator." Previous works using the smoke term methodology suggest that the CC score proposed by Fan, Gordon and Pathak [15] performs well for generating initial candidate smoke terms [1,19]. The CC score algorithm assesses the "relevance" of terms based on how often they occur in relevant versus irrelevant reviews; for example, a term that occurs frequently in

Table 3
Tagger reliability descriptive statistics.

Project	Percentage of student overlap unanimous	Percentage agreement with authorities	Cohen's K	Fleiss, Levin and Paik [16] agreement rating	Landis and Koch [27] agreement rating
Compliments	88.0% (1617/1837)	89.3% (125/140)	0.786	Excellent agreement	Substantial agreement
Feature requests	91.1% (8890/9759)	93.0% (186/200)	0.860	Excellent agreement	Almost perfect agreement
Irritators	85.8% (5104/5950)	86.2% (181/210)	0.724	Fair to good agreement	Substantial agreement

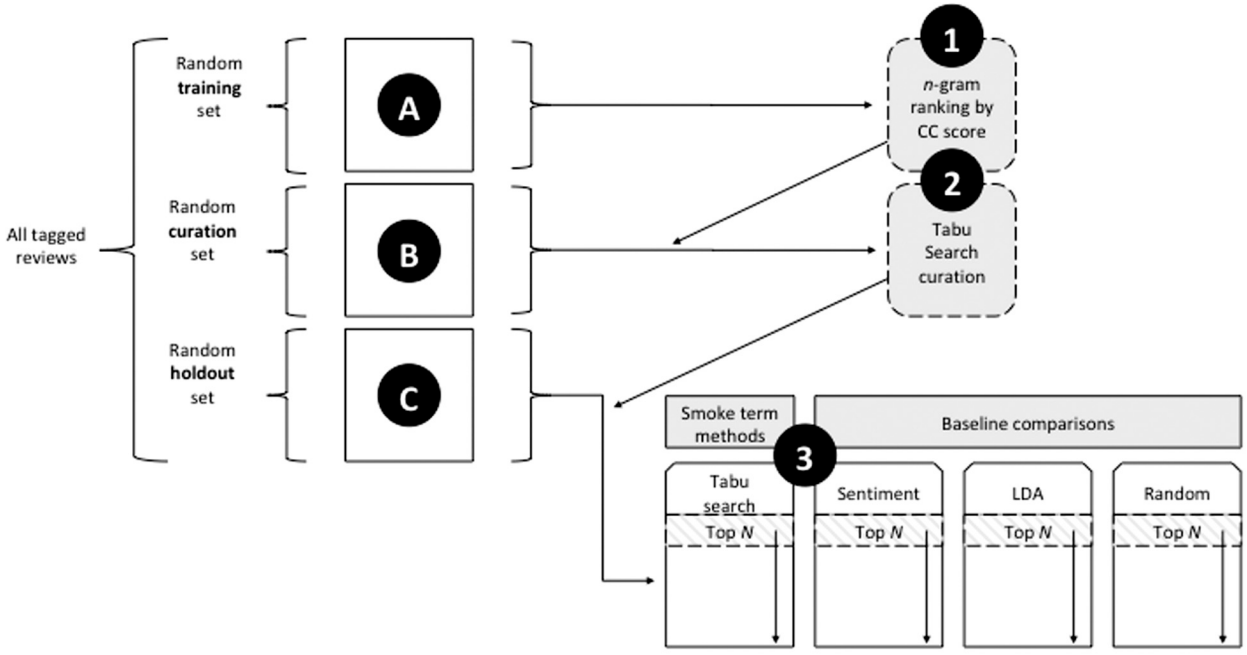


Fig. 1. Schematic of text mining methodology.

"feature request" reviews and is infrequent otherwise would receive a high score [15]. We use the CC score algorithm to generate relevance scores pertaining to all unigrams (one-word), bigrams (two-words), and trigrams (three-words) in our training set (1). These relevance scores are retained, as they serve as weights in a later stage.

Next, we make use of the curation set to refine the initial set of candidate smoke terms (2). For this step, we use the Tabu search heuristic proposed in Goldberg and Abrahams [19]. The authors describe a method by which the algorithm is configured to maximize the precision of potential smoke term lists based on the curation set. That is, if all reviews in the curation set are ranked by a score of the number of instances of each smoke term in each review multiplied by the relative terms' relevance scores (weights), then precision measures the number of reviews in the top *N*-ranked set that refer to true instances. The Tabu search algorithm tests multiple combinations of terms iteratively, each time calculating the precision obtained in the curation set. The algorithm adds terms to the list that improve precision and removes those that harm precision. As the goal of this technique is to choose the top-ranking reviews to read, Goldberg and Abrahams [19] suggest optimizing precision in the top 100-ranked or 200-ranked reviews or using an average of those precision values. Following their example, we maximize the average level of precision in the top 100-ranked and 200-ranked reviews.

The final step in the procedure involves the incorporation of the holdout set, which is unseen by the algorithm thus far. Using the terms generated by the Tabu search heuristic, we test how the selected terms perform given this unseen data using the aforementioned scoring procedure the rank the reviews from most likely to least likely to be of interest to innovation processes (3). The step of testing the efficacy of the terms on unseen data helps to ensure that the terms are not "overfit" and that they generalize well to new datasets [1,19].

4.3. Competing text mining techniques

To ensure the efficacy of our results using the aforementioned text mining techniques, we take the further step of comparing our methods to several alternative text mining methods. The first such method that we consider is sentiment analysis, which is used to assess the emotive content in text. Sentiment analysis is widely applicable in many domains, such as predicting the results of political elections [8] or detecting safety hazard discussions in online media [1]. We use two existing sentiment techniques, AFINN [41] and the Harvard General Inquirer [25] to rank the online reviews in our dataset by their emotive content, and we compare this ranking to our technique. Each sentiment analysis technique assesses text on multiple emotive dimensions, such as positivity, negativity, strength, passivity, pleasure, pain, etc. In this case, we use positivity to assess positive sentiment (compliments) and negativity to assess negative sentiment (irritators).

Additionally, Latent Dirichlet Allocation (LDA), first proposed by Blei et al. [5], is a popular text mining method for mining "topics" in online reviews. The theory underlying LDA suggests that each text corpus is made up of a mixture of several topics. Each document in that corpus (for example, an online review) may contain only some of those topics. Some documents may discuss only one topic, while others may discuss a blend between several topics. LDA is an unsupervised machine learning technique that uses a three-level hierarchical Bayesian model to estimate which terms comprise which topics. LDA has many applications, such as conducting analyses of online chatter about brands in online media [55] and characterizing the topics that make up consumer discussions in service industries [21,58]. We use LDA to determine lists of topics reflecting each attribute type and the extent to which these topics are predictive of innovation opportunities.

5. Results

5.1. Tagging results

After reconciling the tags completed by the student and authority taggers, we first present a description of the tagging results in Table 4. Compliments, feature requests, and irritators were all well-represented in online reviews. Compliments were the best represented at 32.9% of our dataset, whereas irritators constituted 18.0% of the dataset, and feature requests constituted 5.4% of the dataset (note that these figures are not additive as a single review could be coded as referring to multiple attributes). Compliments were most prevalent in reviews that scored high star ratings; however, even 1-star and 2-star reviews were well represented, as even generally critical reviews sometimes note some positive attribute of a product. For example, a consumer may be

Table 4
Star rating distribution for compliments, feature requests, and irritators.

	All firms			Collaborating firm		
Star rating	Compliments	Feature requests	Irritators	Compliments	Feature requests	Irritators
1	800	163	1237	20	18	44
2	646	142	887	9	16	25
3	1277	222	858	24	21	29
4	2340	486	932	285	42	65
5	3176	341	593	594	29	191
Total	8239 (32.9% of reviews)	1354 (5.4% of reviews)	4507 (18.0% of reviews)	932 (42.1% of reviews)	126 (5.7% of reviews)	354 (16.0% of reviews)

dissatisfied with a product or find that it was not a fit for their use case but nonetheless note some positive feature that caused them to purchase it in the first place. Irritators were most associated with negative reviews, although even some high-scoring reviews noted some irritations with otherwise satisfactory products. Feature requests had the weakest association with star ratings; they were slightly more prevalent in high-scoring reviews, but they were most common with 4-star reviews. We also found evidence that these proportions vary by firm: the firm that collaborated with us had far more compliments, slightly more feature requests, and slightly fewer irritators than average. Interestingly, as the firm generally received very high star ratings, most irritators were actually found in 5-star reviews. Therefore, simply analyzing reviews with extreme star ratings is a risky strategy, as each attribute type appears across the full continuum of star ratings.

Our findings suggest that online reviews are a viable medium for discovering innovation opportunities. Each attribute type identified in the adapted attribute mapping framework (see Table 3) is present in online reviews and was reliably coded. However, each attribute is present only in a minority of reviews, and given the great volume of online content, prioritizing the content with text mining techniques to access the most relevant feedback is crucial.

5.2. Text mining results

We employed the aforementioned heuristic text mining method proposed by Goldberg and Abrahams [19] in order to generate unigram, bigram, and trigram terms for each attribute of interest. We display the top five unigrams, bigrams, and trigrams generated by this technique for each attribute in Table 5. Like the findings of prior work, the terms seem to reflect domain-specific jargon [1] as well as narrative structure [19]. Several of the terms, such as "lid," "the alarm," "the unit," or "the top," have particular meanings in the countertop appliances industry in that they refer to specific attributes or components of a product. Even though these terms were all instances of feature requests or irritators, indicating that the consumer was noting some area of improvement for the product, none of these terms explicitly states a negative experience. These words and phrases are unlikely to be well recognized by sentiment approaches, which are not tuned to the specific nuances of the countertop appliances domain. In the context of this domain, however, consumers only tend to use these terms when they are describing an issue with some component of the product. An irritator might complain "I couldn't get **the top** to stay on," and this type of usage, which is largely non-emotive, is hugely prevalent in domain-specific WOM [19]. In addition, many of the terms do not refer specifically to a product attribute, but they instead identify the narrative in which the customer describes their experience. For example, the phrase "stars is because" reflects instances in which the reviewer attempts to justify the star rating in their review by noting some experience with an aspect of the product that affected their final rating. Reviewers using this phrase do so with the expectation that others will read their review, and they preemptively justify their star rating to those readers.

Table 5
Top terms generated by the Tabu search heuristic [19].

Panel A: Top unigram, bigram, and trigram compliment terms.					
Unigram	Weight	Bigram	Weight	Trigram	Weight
Easy	96,308	easy to	85,383	easy to clean	63,277
Highly	36,742	very easy	39,662	easy to use	53,435
Durable	19,122	so easy	26,518	i love this	26,117
Fantastic	17,462	fast and	17,855	so easy to	21,900
Owned	16,170	clean i	15,247	highly recommend it	12,500
Panel B: Top unigram, bigram, and trigram feature request terms					
Unigram	Weight	Bigram	Weight	Trigram	Weight
Wish	33,510	wish it	28,360	i wish it	19,662
Needs	14,100	wish the	19,993	have been nice	15,945
Perhaps	13,878	would be	17,176	stars is because	14,598
Lid	12,506	been nice	15,945	could have been	12,929
Change	10,319	the alarm	13,934	if it had	12,929
Panel C: Top unigram, bigram, and trigram irritator terms					
Unigram	Weight	Bigram	Weight	Trigram	Weight
Return	49,499	would not	40,174	do not buy	36,444
Disappointed	38,754	does not	38,513	not buy this	31,392
First	38,406	the unit	36,455	i have to	29,546
Stopped	31,745	not recommend	35,266	piece of junk	29,385
Way	30,577	the top	29,473	would not recommend	28,642

We also note that while some of the terms above do not indicate the target classes themselves, we may still see them occur alongside the target class. For example, we identified "return" as an irritator. A product return indicates that the customer was so unsatisfied with a product that they changed their purchasing decision and potentially made a further purchasing decision to choose a competing product. The negative purchasing decision is indicative of an irritator, but the word "return" does not diagnose the cause of the irritator in itself. However, identifying reviews that contain "return" is useful in our context because it suggests that those reviews contain additional information that is explanatory, such as "I had to **return** the product because **the top** wouldn't stay on."

Using the unseen holdout set, we compare our technique to several other text mining techniques to benchmark its performance. In addition to sentiment analyses [25,41], and we also used LDA [5] to generate topics that may be predictive of the attributes from the attribute mapping framework. We ran the LDA algorithm for 1500 iterations and generated 10 topics of 15 words each, displayed in Table 6.[1] We manually labeled each topic based on its contents [21]. Most topics identified different types of products, but we identified topic #2 as denoting negative product experiences, which we used to predict irritators, and topic #4 as denoting positive product experiences, which we used to predict compliments. None of the topics seemed to relate to feature requests.

We use each of these methods to rank the reviews in our dataset from

[1] We tested running LDA for various numbers of topics ranging from 2 to 25. When fewer topics were identified, none seemed to refer to positive or negative product experiences. When more topics were identified, LDA generated more topics that pertained to types of countertop appliances or redundant topics.

Table 6
10-topic analysis output from LDA [5].

Topic	Top terms
#1: Pans and cookware	pan, set, pans, stick, use, it, non, cooking, cook, cookware, great, heat, well, clean, stainless
#2: Negative product experience	one, product, amazon, back, it, unit, get, first, reviews, new, could, time, buy, made, replacement
#3: Water filters	water, air, kettle, it, filter, unit, room, use, filters, smell, much, really, dust, clean, fan
#4: Positive product experience	it, great, easy, use, love, product, one, works, well, recommend, bought, price, opener, clean, gift
#5: Purchasing narrative	one, years, it, bought, needed, old, used, new, last, year great, use, another, model, still
#6: Blenders/juicers	blender, ice, it, use, make, cream, clean, juicer, machine, great, easy, juice, blade, food, get
#7: Slow/rice cookers, mixers	rice, mixer, cooker, it, pot, use, time, one, cooking, bowl, cook, slow, great, love, used
#8: Coffee makers	coffee, cup, maker, water, it, machine, hot, pot, use, carafe grinder, one, great, brew, makes
#9: Popcorn and waffle makers	popcorn, easy, it, use, pop, great, waffle, clean, cooking, one, time, cook, make, oil, waffles
#10: Toasters	toaster, it, oven, toast, lid, top, one, use, well, unit, makes, get, time, bread, nice

most likely to least likely to contain a compliment, feature request, and/or irritator. Assessing the efficacy of these techniques involves choosing an arbitrary cutoff (the top *N*-ranked reviews for a given technique) and computing the number or percentage of true instances of each attribute within that cutoff. In Table 7, we show the performance of each technique at the top 200 reviews, following the example of previous works [1,19], as this volume of content might be considered manageable for many firms.[2] The top performing technique for each attribute is indicated in **bold**. For each of compliments, feature requests, and irritators, our domain-specific terms far outperform the competing techniques. LDA was the nearest performing alternative for compliments, but it was the worst performing alternative for irritators. Simple star ratings bested both LDA and sentiment analysis at predicting irritators, but they were very poor at predicting feature requests. For each attribute, we compared our domain-specific techniques to the competing techniques using a chi-squared test; each technique significantly differed from each competing technique at the 0.001 level. In Figs. 2, 3, and 4, we display the performance of each technique over a range of possible cutoffs.

Table 7
Performance of each technique within the top 200-ranked reviews.

	Number (percentage) of true instances in top 200-ranked reviews		
Technique	Compliments	Feature requests	Irritators
Unigrams	171 (85.5%)	62 (31.0%)	**148 (74.0%)**
Bigrams	**173 (86.5%)**	87 (43.5%)	140 (70.0%)
Trigrams	**173 (86.5%)**	**103 (51.5%)**	135 (67.5%)
LDA	116 (58.0%)	–	67 (33.5%)
AFINN	95 (47.5%)	27 (13.5%)	77 (38.5%)
Harvard GI	85 (42.5%)	31 (15.5%)	63 (31.5%)
Star ratings	102 (51.0%)	8 (4.0%)	97 (48.5%)

[2] We assume a context in which industry professionals do not have the time to analyze every possible review, but only a maximally useful subset. For robustness, we experimented with other supervised learning approaches, utilizing information retrieval techniques such as TF-IDF and deep learning techniques such as BERT to perform classifications. These classifiers tend to be very accurate over the entirety of the dataset; however, in our use case focused on just the top-ranked shortlist of reviews, they were less accurate. For instance, our highest-performing technique, BERT, was 80% accurate in the top 200-ranked compliment reviews. Thus, for brevity, we focus on the methodologies optimized for the ranking-based approach.

5.3. Case study

In the following, we show a short case study for deploying automated detection of innovation opportunities in online reviews. We chose a line of coffee makers by a competitor of the collaborating countertop appliance manufacturer for our case study. We filtered the reviews pertaining to those products, and we ranked those reviews based on the domain-specific terms generated previously. In Table 8, we show a selection of top-scoring reviews for each of these attribute types. Terms in the domain-specific dictionaries are indicated in **bold**, and specific feedback for the firm's product offerings is underlined. As previous work has suggested, the reviews tend to follow a narrative structure in which the reviewer details their experience with the product [19]. Each of top-ranking compliment reviews praised the product's ease of use, and each top-ranking feature request review requested that the product be redesigned to handle different or larger cup sizes, a concern shared by the top compliment review. The second feature request review advises readers to purchase competing products on that basis. The top two irritator reviews both express frustration at the product's leaking, possible due to a faulty gasket, while the third irritator review complains that the brews made by their machine have shrunken over time. The first and third irritator reviews indicate that the irritation affects their purchasing decisions, as they otherwise enjoy the product but would opt for an alternative if the problem persists.

These reviews present clear instances in which consumer purchasing decisions are directly related to product attributes. Using the rapid feedback from a small sampling of prioritized reviews, the firm can immediately interpret their position in the context of the revised attribute mapping framework, shown in Table 9. The issues with brew sizes and the leaking/faulty gasket may sway many consumers, including those reading the online reviews, to purchase alternative products. The firm may prioritize these fixes in their product development process. Meanwhile, the firm can market its product as being easy to use, which makes a positive impression on their customers.

A competing methodology to our approach described above is to utilize aspect mining to identify potential areas of strength or weakness in a product. While there are several possible methodologies, one common thread is that unsupervised topic modeling, such as LDA, is used to identify a series of topics discussed in a corpus, and then sentiment analysis is used to evaluate the relative emotive content for each topic [3,45]. For example, in our context, we might observe a topic for gaskets with a particularly negative sentiment, a topic for ease of use with particularly positive sentiment, etc.

To evaluate the efficacy of this alternative in our case study, we implement an aspect mining methodology. First, we utilized LDA to identify the predominant topics among the line of coffee makers, as shown in Table 10 below. As Debortoli, Müller, Junglas and vom Brocke [12] note, the choice in number of topics chosen for LDA vitally affects results; per the authors' suggestion, we experimented with various numbers of topics and chose five topics, as this was more easily human-interpretable than other options. Based on the methodology of Srinivas and Rajendran [53], we then computed the relative sentiment for reviews referring to those topics to assign a consolidated sentiment score to each topic. We scaled these values such that +1 reflects the most positive sentiment, and − 1 reflects the most negative sentiment. For brevity, these consolidated sentiment scores were based on AFINN, but we also observed similar sentiment scores when experimenting with other sentiment analysis methods. We also show this information in Table 10 below.

Results of the aspect mining analysis interestingly differ from the smoke term-based analysis. As the aspect mining analysis is driven by LDA's unsupervised topic determinations, many of the topics, such as topics 1 and 2, are relatively high-level and vague. In some cases, LDA-derived topics may be consistent with human judgement, but in other cases, they may merely satisfy algorithmic fit. As most reviews in the dataset were generally positive (most reviews reflect high star ratings

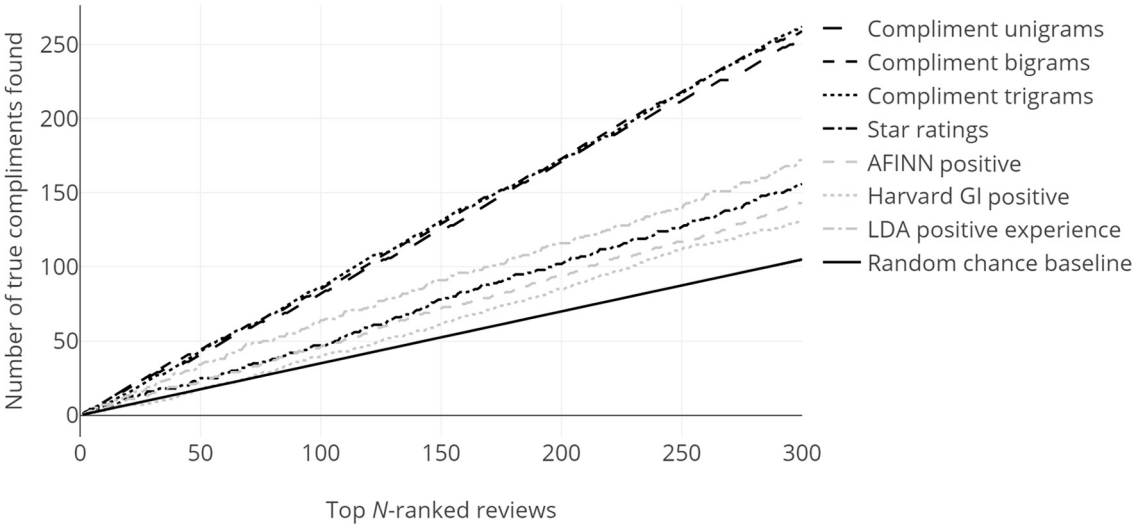


Fig. 2. Lift chart of text mining performance predicting compliments.

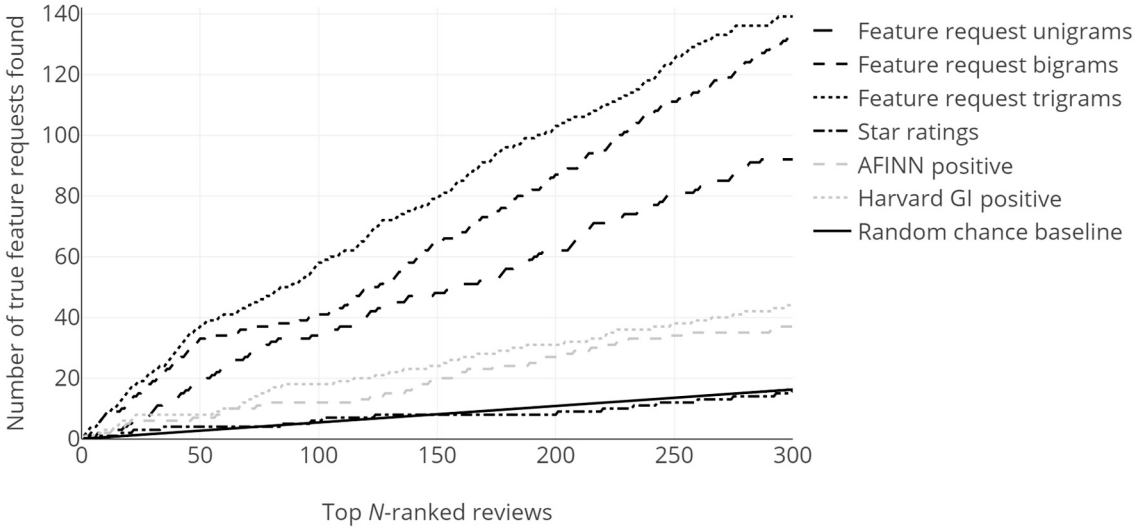


Fig. 3. Lift chart of text mining performance predicting feature requests.

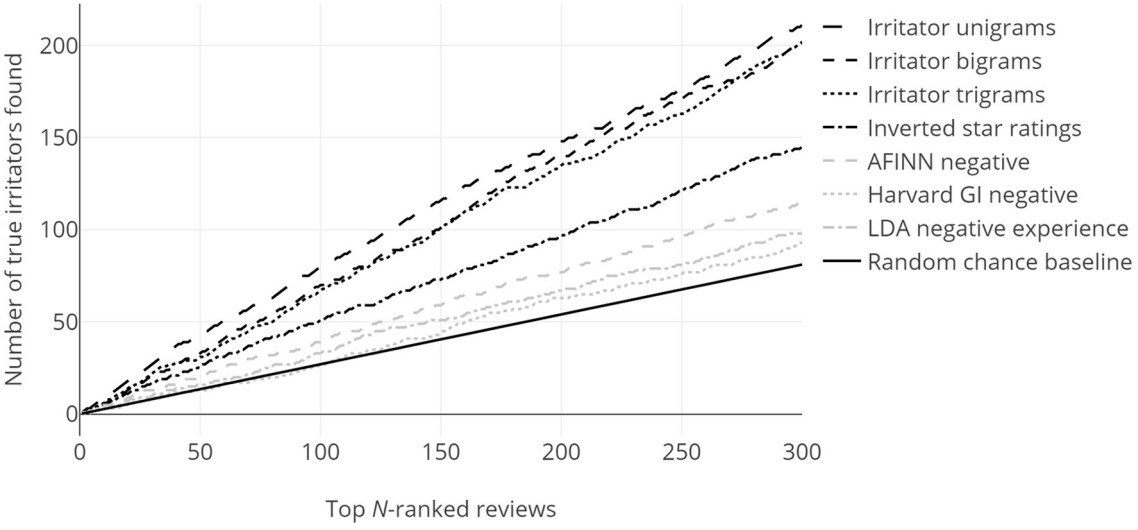


Fig. 4. Lift chart of text mining performance predicting irritators.

Table 8
Top-scoring reviews by attribute type for coffee makers.

Top compliments	Top feature requests	Top irritators
Overall **I love this** machine. It looks sleek; very **easy to use**. My only complaint is the amount of water that is required to be kept in the reservoir even for a small amount of coffee.	I like my [brand name]. It makes a consistent cup of coffee. This was the cheaper version of the different [brand name] machines available and Im very glad that I got it. **I do wish there** was a way to **change** the amount of coffee brewed. There is I suppose, but that would mean a more expensive machine. It makes about half a cup of what I would consider a normal cup of coffee. Other than that I am very happy with the machine…	… overflows water and grounds into the receiving cup due to a faulty gasket above the upper needle. No instructions tell you that **you have to** lower the gasket ring above the needle each time you use the [brand name] - forget and the coffee is not fit to drink, lower it too far and it comes off. This is not an improvement. This is a great item if it worked as advertised. I **would not recommend** it at this time.
This this is amazing. I honestly cannot say one bad thing about it. Extremely **easy to use,** great coffee (of course that also depends on what kind of coffee you use), and perfect size for my desk.	This unit works great. It's simple and functional. However, if you're purchasing something for your home, opt for a better model or you will be disappointed. This is **functional** and small for an office or an area away from home but for the home, trust me, you want the model **with different** cup sizes and the water reservoir because it makes the experience so much better…	We loved our [brand name], for about 18 months. Then it started **leaking** from the bottom. We tried adjusting the gasket where the needle comes in per several suggestions. then we took it apart and tried to clean and adjust a gasket inside. Still **leaks**…
I bought this for my daughter, she loves it **easy to use** and convenient, bought some of the pods that can be used with regular coffee and saves some money	I bought this to use at work and I love it. I wanted something a little bit compact to use in my office. The only negitive is that it takes 3 minutes to get my coffee. **I do wish it** would hold a few cups of water so **it would be** more instant like the one I have at home.	like other reviews, we have found that the more we use this unit, the smaller our cups of coffee get. not sure what the issue is. i am hoping to find a fix or that the cup size will stop shrinking. if not i will **return the unit**. other than that issue, it works great and is perfect for our house, where i like tea and my wife likes coffee.

Table 9
Attribute map for coffee makers based on online review intelligence.

	Basic	Discriminators	Energizers
Positive	Non-negotiables	Compliments	
		Ease of use	
Negative	Tolerables	Feature requests	
		Different/larger brew sizes	
		Irritators	
		Brew size decreasing over time	
		Leaking/faulty gasket	
Neutral	So whats?	Parallel differentiators	N/A

and positive sentiment), all consolidated sentiment scores ranged from 0.53 to 0.90. Thus, none of the insights garnered from the smoke term-based analysis, such as the compliments about ease of use, the irritators about brew size and faulty gaskets, and the feature requests about different brew sizes would be discovered.

5.4. Validation of usefulness

Authority taggers were asked to offer their opinion on the usefulness of each review that they tagged in a three-point scale. In doing so, we enable a comparison of the usefulness of each attribute type (compliments, feature requests, and irritators) versus other online reviews. Table 11 presents the authority taggers' tagging counts for each of these attribute types. In each attribute type, the authority taggers indicated that they found the target classification reviews more useful than alternative reviews that they were provided. As our data is ordinal and non-normal, we assess the difference between each attribute and alternative reviews statistically using a Mann-Whitney *U* test [36]. These statistical tests indicated that tags for each attribute type, compliments, feature requests, and irritators, differed significantly from alternative reviews at the 0.001 level. This statistical evidence suggests that online reviews are a meaningful source of innovation ideas that are beneficial to product development teams, and it provides an empirical basis that the attribute mapping framework applies in this domain [34,35]. One potential source of bias in the authority taggers' assessments of usefulness is since they tagged for attribute mapping components and usefulness at the same time. For robustness against this potential cross-contamination, we used a holdout authority tagger from the collaborating countertop appliance manufacturer for one further round of tagging. We presented this tagger with a stratified random sample of reviews in which 1/6 had been identified as compliments, 1/6 had been identified as feature requests, 1/6 have been identified as irritators, and the remaining 1/2 did not fall into any attribute type of interest. The holdout tagger only tagged for usefulness on the three-point scale without regard for the attribute mapping constructs. This tagger's results are displayed in Table 12. Like the other authority taggers, this tagger rated compliments, feature requests, and irritators as more useful than alternative reviews, and their tags in each attribute type significantly differed from the alternative reviews at the 0.001 level.

Table 10
Aspect mining analysis of coffee maker reviews.

Topic	Key terms	Consolidated sentiment score
1: Coffee maker	Machine, reservoir, clean, noise, setup	+0.79
2: Coffee/tea	Tea, coffee, espresso, bean, selection	+0.90
3: Product functionality	Cup, make, oz., pot, brewing	+0.82
4: Time	Time, month, morning, day, work	+0.66
5: Product size	Mini, thing, plastic, bigger, size	+0.53

6. Conclusions

This paper presents the first large-scale empirical validation of the popular attribute mapping framework by MacMillan and McGrath [34,35]; as opposed to case studies, we find enormous statistical evidence not only that consumers post feedback that aligns with the attribute mapping framework, but also that firms can glean valuable insights from that feedback. This empirical validation lends credence to many other works that incorporate this theoretical framework [2,38,52]. Our text mining results suggest that the exaptation of the Goldberg and Abrahams [19] methodology applies to product innovation opportunities in online reviews, responding to the call by Lee and Bradlow [29] for the development of data mining technologies that directly consider user needs. Particularly since hastened product and business life cycles have increased innovation pressures on firms [42], the capability to rapidly source innovation-related feedback from online media is imperative. Many potential applications exist for practitioners in these rapid prioritization technologies. Most obviously, the proposed technique allows for firms to source innovation-related feedback quickly and easily from a mass of consumers and map it into a verified framework for organizing and prioritizing product attributes. In addition to their own

Table 11
Authority taggers' perceptions of online review usefulness.

	Number (percentage) of authority tags by attribute type and usefulness					
	Compliments		Feature requests		Irritators	
Usefulness	Compliment	No	Feature request	No	Irritator	No
Not useful at all	7 (10.6%)	39 (52.7%)	36 (36.0%)	71 (71.0%)	22 (20.8%)	62 (59.6%)
A bit useful	35 (53.0%)	30 (40.5%)	46 (46.0%)	25 (25.0%)	62 (58.5%)	34 (32.7%)
Very useful	24 (36.4%)	5 (6.8%)	18 (18.0%)	4 (4.0%)	22 (20.8%)	8 (7.7%)
Total	66	74	100	100	106	104

Table 12
Holdout tagger's perceptions of online review usefulness.

	Compliment	Feature request	Irritator	None
Not useful at all	10 (47.6%)	4 (20.0%)	4 (19.0%)	56 (82.4%)
A bit useful	8 (38.1%)	8 (40.0%)	13 (61.9%)	11 (16.1%)
Very useful	3 (14.2%)	8 (40.0%)	4 (19.0%)	1 (1.4%)
Total	21	20	21	68

products, firms can perform analyses of competing products to discern the source of competitive advantages and thereby to obtain a superior position. This analysis can be performed prospectively, searching for new insights using customer-driven feedback, or retrospectively, using consumer-driven feedback to verify or update an existing perception.

Our work is subject to several limitations. First, we relied on a large team of volunteer taggers to help code the data used in this study, and our supervised learning technique relied on these volunteers' opinions, which introduces a source of bias and variability. Delineating between one attribute type and another (e.g., "feature request" versus "no feature request") is a somewhat subjective process, as it relies upon each individual's reading of the online review. We used several safeguards in this work to minimize this variability as much as possible, including providing the taggers with a detailed protocol document and comparing the tags to authorities, with whom they had considerable agreement. Second, our techniques should serve as a useful component of innovation efforts, but they should not be used alone. Prior work has extolled the virtues of many sources for innovation ideas, including brainstorming, focus groups, consumer surveys, warranty claims, and competitive monitoring [44]. Our technique helps firms to harness a massive volume of online consumer-driven data, but it does not serve to replace existing methods. Another limitation to our work is that we consolidated some parts of the original attribute mapping framework for our analyses. While we felt this was appropriate for our work, we acknowledge that it would be useful in practice to delineate dissatisfiers from enragers (consolidated as "irritators") and to delineate differentiators from exciters (consolidated as "compliments"). Our approach did not allow for this level of granularity; however, our approach could be used in combination with more traditional market research to derive further detail. For example, focus groups could be used to identify how broadly a concern about a leaking/faulty gasket is salient; then, this irritator could be reclassified as a dissatisfier or enrager.

In future work, in addition to prioritizing online reviews using the framework described in this paper, practitioners may also like to aggregate reviews of similar topics together to understand overarching trends. After our technique is run as an initial stage, these techniques may, for example, allow firms to allocate reviews to relevant teams that can directly address innovation opportunities. This step is beyond the scope of this paper, but a topic mining technique such as LDA [5] or a simpler bag-of-words model may assist in this process. Another possible extension of this work concerns extending our framework to other forms of online media, such as social media, news, and/or forum posts. Online reviews are clearly associated with specific products, so it is clear which reviews pertain to relevant products; on the broader web, further techniques could assist in rapidly sifting through different forms of textual data and identifying pressing innovation-related content.

Author statement

D.M.G and A.S.A performed and supervised data collection and proposed the study. D.M.G performed statistical and text analytic analyses and wrote the paper. D.M.G and A.S.A edited the paper. D.M.G and A.S.A contributed to the theoretical development and direction of the paper. The authors are grateful to the Apex Systems Center for Innovation and Entrepreneurship for supporting this research.

References

[1] A.S. Abrahams, W. Fan, G.A. Wang, Z.J. Zhang, J. Jiao, An integrated text analytic framework for product defect discovery, Prod. Oper. Manag. 24 (6) (2015) 975–990.
[2] R. Amit, C. Zott, Value creation in e-business, Strateg. Manag. J. 22 (6–7) (2001) 493–520.
[3] R.K. Amplayo, S. Lee, M. Song, Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis, Inf. Sci. 454 (2018) 200–215.
[4] L.A. Bettencourt, A.W. Ulwick, The customer-centered innovation map, Harv. Bus. Rev. 86 (5) (2008) 109.
[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[6] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.
[7] BrightLocal, Local Consumer Review Survey 2016, in, 2016.
[8] A. Ceron, L. Curini, S.M. Iacus, G. Porro, Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France, New Media Soc. 16 (2) (2014) 340–358.
[9] M.-H. Chang, J.E. Harrington Jr., Innovators, imitators, and the evolving architecture of problem-solving networks, Organ. Sci. 18 (4) (2007) 648–666.
[10] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: online book reviews, J. Mark. Res. 43 (3) (2006) 345–354.
[11] F. Damanpour, J.D. Wischnevsky, Research on innovation in organizations: distinguishing innovation-generating from innovation-adopting organizations, J. Eng. Technol. Manag. 23 (4) (2006) 269–291.
[12] S. Debortoli, O. Müller, I.A. Junglas, J. vom Brocke, Text mining for information systems researchers: an annotated topic modeling tutorial, Commun. AIS 39 (2016) 7.
[13] K.C. Desouza, Y. Awazu, S. Jha, C. Dombrowski, S. Papagari, P. Baloh, J.Y. Kim, Customer-driven innovation, Res. Technol. Manag. 51 (3) (2008) 35–44.
[14] M. du Plessis, The role of knowledge management in innovation, J. Knowl. Manag. 11 (4) (2007) 20–29.
[15] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decis. Support. Syst. 40 (2) (2005) 213–233.
[16] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, John Wiley & Sons, 2013.
[17] D.C. Galunic, S. Rodan, Resource recombinations in the firm: knowledge structures and the potential for Schumpeterian innovation, Strateg. Manag. J. (1998) 1193–1201.
[18] M. Ghiassi, D. Zimbra, S. Lee, Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks, J. Manag. Inf. Syst. 33 (4) (2016) 1034–1058.
[19] D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decis. Support. Syst. 105 (2018) 52–65.
[20] D.M. Goldberg, S. Khan, N. Zaman, R.J. Gruss, A.S. Abrahams, Text mining approaches for postmarket food safety surveillance using online media, Risk Anal. (2020). In press.
[21] Y. Guo, S.J. Barnes, Q. Jia, Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation, Tour. Manag. 59 (2017) 467–483.
[22] N. Hu, J. Zhang, P.A. Pavlou, Overcoming the J-shaped distribution of product reviews, Commun. ACM 52 (10) (2009) 144–147.

[23] C. Iacob, R. Harrison, Retrieving and analyzing mobile apps feature requests from online reviews, in: Proceedings of the 10th Working Conference on Mining Software Repositories, IEEE Press, 2013, pp. 41–44.
[24] P.W. Jackson, S. Messick, The person, the product, and the response: conceptual problems in the assessment of creativity, J. Pers. 33 (3) (1965) 309–329.
[25] E.F. Kelly, P.J. Stone, Computer Recognition of English Word Senses, North-Holland, 1975.
[26] B. Kim, J. Park, J. Suh, Transparency and accountability in AI decision support: explaining and visualizing convolutional neural networks for text information, Decis. Support. Syst. 113302 (2020).
[27] J.R. Landis, G.G. Koch, The Measurement of Observer Agreement for Categorical Data, Biometrics, 1977, pp. 159–174.
[28] D. Law, R. Gruss, A.S. Abrahams, Automated defect discovery for dishwasher appliances from online consumer reviews, Expert Syst. Appl. 67 (2017) 84–94.
[29] T.Y. Lee, E.T. Bradlow, Automated marketing research using online customer reviews, J. Mark. Res. 48 (5) (2011) 881–894.
[30] J.J. Li, X.-P. Chen, S. Kotha, G. Fisher, Catching fire and spreading it: a glimpse into displayed entrepreneurial passion in crowdfunding campaigns, J. Appl. Psychol. 102 (7) (2017) 1075.
[31] X. Li, L.M. Hitt, Self-selection and information role of online product reviews, Inf. Syst. Res. 19 (4) (2008) 456–474.
[32] J.C. Linder, S. Jarvenpaa, T.H. Davenport, Toward an innovation sourcing strategy, MIT Sloan Manag. Rev. 44 (4) (2003) 43–50.
[33] G.T. Lumpkin, G.G. Dess, Clarifying the entrepreneurial orientation construct and linking it to performance, Acad. Manag. Rev. 21 (1) (1996) 135–172.
[34] I.C. MacMillan, R.G. McGrath, Discover your products' hidden potential, Harv. Bus. Rev. 74 (3) (1996) 58–73.
[35] I.C. MacMillan, R.G. McGrath, Discovering new points of differentiation, Harv. Bus. Rev. 75 (1997) 133–145.
[36] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. (1947) 50–60.
[37] J. McAuley, R. Pandey, J. Leskovec, Inferring networks of substitutable and complementary products, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 785–794.
[38] R.G. McGrath, A. Nerkar, Real options reasoning and a new look at the R&D investment strategies of pharmaceutical firms, Strateg. Manag. J. 25 (1) (2004) 1–21.
[39] D.L. Meadows, Estimate accuracy and project selection models in industrial research, Industr. Manag. Rev. 9 (3) (1968) 105.
[40] V. Mummalaneni, R. Gruss, D.M. Goldberg, J.P. Ehsani, A.S. Abrahams, Social media analytics for quality surveillance and safety hazard detection in baby cribs, Saf. Sci. 104 (2018) 260–268.
[41] F.Å. Nielsen, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, arXiv (2011) preprint arXiv: 1103.2903.
[42] A. Pérez-Luño, J. Wiklund, R.V. Cabrera, The dual nature of innovative activity: how entrepreneurial orientation influences innovation generation and adoption, J. Bus. Ventur. 26 (5) (2011) 555–571.
[43] D. Pessach, G. Singer, D. Avrahami, H.C. Ben-Gal, E. Shmueli, I. Ben-Gal, Employees recruitment: a prescriptive analytics approach via machine learning and mathematical programming, Decis. Support. Syst. 113290 (2020).
[44] J. Pruitt, T. Adlin, The Persona Lifecycle: Keeping People in Mind throughout Product Design, Elsevier, 2010.
[45] J. Qi, Z. Zhang, S. Jeon, Y. Zhou, Mining customer requirements from online reviews: a product improvement perspective, Inf. Manag. 53 (8) (2016) 951–963.
[46] Z. Qiao, G.A. Wang, M. Zhou, W. Fan, The impact of customer reviews on product innovation: empirical evidence in mobile apps, in: Analytics and Data Science, Springer, 2018, pp. 95–110.
[47] N. Rosenbusch, J. Brinckmann, A. Bausch, Is innovation always beneficial? A meta-analysis of the relationship between innovation and performance in SMEs, J. Bus. Ventur. 26 (4) (2011) 441–457.
[48] S. Sachan, J.-B. Yang, D.-L. Xu, D.E. Benavides, Y. Li, An explainable AI decision-support-system to automate loan underwriting, Expert Syst. Appl. 144 (2020), 113100.
[49] H. Sarooghi, D. Libaers, A. Burkemper, Examining the relationship between creativity and innovation: a meta-analysis of organizational, cultural, and environmental factors, J. Bus. Ventur. 30 (5) (2015) 714–731.
[50] R. Sethi, D.C. Smith, C.W. Park, Cross-functional product development teams, creativity, and the innovativeness of new consumer products, J. Mark. Res. 38 (1) (2001) 73–85.
[51] M. Siering, A.V. Deokar, C. Janze, Disentangling consumer recommendations: explaining and predicting airline recommendations based on online reviews, Decis. Support. Syst. 107 (2018) 52–63.
[52] D.G. Sirmon, M.A. Hitt, Managing resources: linking unique resources, management, and wealth creation in family firms, Entrepr. Theory Pract. 27 (4) (2003) 339–358.
[53] S. Srinivas, S. Rajendran, Topic-based knowledge mining of online student reviews for strategic planning in universities, Comput. Ind. Eng. 128 (2019) 974–984.
[54] S. Stieglitz, L. Dang-Xuan, Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior, J. Manag. Inf. Syst. 29 (4) (2013) 217–248.
[55] S. Tirunillai, G.J. Tellis, Mining marketing meaning from online chatter: strategic brand analysis of big data using latent dirichlet allocation, J. Mark. Res. 51 (4) (2014) 463–479.
[56] R.W. Vossen, Relative strengths and weaknesses of small firms in innovation, Int. Small Bus. J. 16 (3) (1998) 88–94.
[57] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decis. Support. Syst. 90 (2016) 23–32.
[58] Z. Xiang, Q. Du, Y. Ma, W. Fan, A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism, Tour. Manag. 58 (2017) 51–65.
[59] K. Xu, S.S. Liao, J. Li, Y. Song, Mining comparative opinions from customer reviews for competitive intelligence, Decis. Support. Syst. 50 (4) (2011) 743–754.
[60] N. Zaman, D.M. Goldberg, A.S. Abrahams, R.A. Essig, Facebook hospital reviews: automated service quality detection and relationships with patient satisfaction, Decis. Sci. 52 (6) (2021) 1403–1431.
[61] N. Zaman, D.M. Goldberg, R.J. Gruss, A.S. Abrahams, S. Srisawas, P. Ractham, M. M. Şeref, Cross-category defect discovery from online reviews: supplementing sentiment with category-specific semantics, Inf. Syst. Front. (2021) 1–21.
[62] H. Zhang, H. Rao, J. Feng, Product innovation based on online review data mining: a case study of Huawei phones, Electron. Commer. Res. 18 (1) (2018) 3–22.
[63] M. Zhang, B. Fan, N. Zhang, W. Wang, W. Fan, Mining product innovation ideas from online reviews, Inf. Process. Manag. 58 (1) (2021), 102389.

David M. Goldberg is Assistant Professor of Management Information Systems in the Fowler College of Business at San Diego State University. He received his doctoral and bachelor's degrees from Virginia Tech. His current research interests are in the areas of text mining, machine learning, decision support systems, and expert systems. He has published in *Decision Support Systems*, *Decision Sciences, Risk Analysis*, *Information Technology & Management*, *Communications of the Association for Information Systems*, and others.

Alan S. Abrahams is an Associate Professor in the Department of Business Information Technology at Virginia Tech and a member of the Affiliated Faculty at the Center for Injury Research and Policy at the Johns Hopkins Bloomberg School of Public Health. His research on quality analytics from text is published in *Production and Operations Management*, *Decision Support Systems*, and other high-impact journals. He holds a PhD in Computer Science from the University of Cambridge, and a Bachelor of Business Science degree in Information Systems from the University of Cape Town.