BH 722

Business Horizons (2016) 59, 115—124

Available online at www.sciencedirect.com

ScienceDirect

KELLEY SCHOOL OF BUSINESS
INDIANA UNIVERSITY

www.elsevier.com/locate/bushor

Uncovering the message from the mess of big data

Neil T. Bendle*, Xin (Shane) Wang

Ivey Business School, Western University, 1255 Western Road, London, Ontario, N6G 0N1, Canada

KEYWORDS
Big data;
User-Generated content;
Latent Dirichlet Allocation;
Topic modeling;
Market research;
Qualitative data

Abstract User-generated content, such as online product reviews, is a valuable source of consumer insight. Such unstructured big data is generated in real-time, is easily accessed, and contains messages consumers want managers to hear. Analyzing such data has potential to revolutionize market research and competitive analysis, but how can the messages be extracted? How can the vast amount of data be condensed into insights to help steer businesses' strategy? We describe a non-proprietary technique that can be applied by anyone with statistical training. Latent Dirichlet Allocation (LDA) can analyze huge amounts of text and describe the content as focusing on unseen attributes in a specific weighting. For example, a review of a graphic novel might be analyzed to focus 70% on the storyline and 30% on the graphics. Aggregating the content from numerous consumers allows us to understand what is, collectively, on consumers' minds, and from this we can infer what consumers care about. We can even highlight which attributes are seen positively or negatively. The value of this technique extends well beyond the CMO's office as LDA can map the relative strategic positions of competitors where they matter most: in the minds of consumers.

1. Consumers and the world of big data

Understanding what consumers want is a fundamental business problem that is being radically changed by new technology. It used to be that consumers whispered and the challenge facing managers was to try and get them to speak up. Now, consumers shout and the challenge facing managers is to uncover the messages hidden among the crescendo of overlapping voices.

1.1. Listening to the market

Traditional listening techniques (e.g., focus groups, surveys) can be very useful but are typically expensive, are limited in scope, and require great skill to

* Corresponding author
E-mail addresses: nbendle@ivey.ca (N.T. Bendle), xwang@ivey.uwo.ca (X. Wang)

0007-6813/$ — see front matter © 2015 Kelley School of Business, Indiana University. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bushor.2015.10.001

run effectively. Because these exercises are formally scheduled, the voice of the market tends to only emerge in short bursts and infrequently. Even when investing heavily in research, a firm only gets feedback on a modest number of attributes. If the market researcher doesn't ask the right questions, a firm may not uncover what matters to consumers. Finally, there are also significant problems if the consumer finds it hard to fully verbalize the answers when put on the spot.

The world has changed and big data is transforming many elements of business, from analytics (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011) to talent management (Russell & Bennett, 2015). Analysts have uncovered gems of wisdom about their customers by integrating data from different parts of the organization (Thelen, Mottner, & Berman, 2004). Integrating data held in your organization is an excellent way of improving knowledge of your customers but gives a limited picture. Most of the information about your customers isn't held anywhere on the company servers; it is housed on various websites that are typically as visible to your competitors as they are to you.

Furthermore, many of the key insights that managers wish to uncover are about potential customers. These consumers aren't currently purchasing from your firm but could be enticed to do so. When approaching potential customers, your firm doesn't have any proprietary insight into their needs. No crunching of internal data will allow you to better understand what they want. In such a world, insight comes from being able to look outside your organization for information.

The recent proliferation of user-generated content—such as product reviews, tweets, and blogs—has provided numerous ways for consumers to share their opinions. Rather than subscribing to any dystopian vision of firms spying on consumers, we believe the big data revolution can produce firms that better respond to consumers' wishes. Consumers want to be heard. They post about the successes and failures of products, brands, and firms precisely because they feel that their views should be listened to. Well-managed organizations agree and want to listen to what consumers have to say.

In this new world of social media, opinion sites, and comment threads, consumers have, collectively, not been shy about expressing themselves. Your firm's analysts, and analysts working for your competitors, can find out what consumers think about your firm—often in brutally frank detail. One only need look at Amazon.com to understand just how extensive is the trove of data hidden in plain sight. Consumers share their thoughts about every type of product, from the artistic (decorative garden water features) to the fun (children's Halloween costumes) to the insightful (marketing analytics reference books). While all firms can benefit, the new technology presents an especially valuable opportunity regarding consumer marketing for firms involved in high volume business. Historically, the sheer volume of current and potential customers has presented managers with great difficulties in understanding what consumers collectively want. Now, though, the wide range of people who use the product and then share their opinions produces a massive online, real-time comments box.

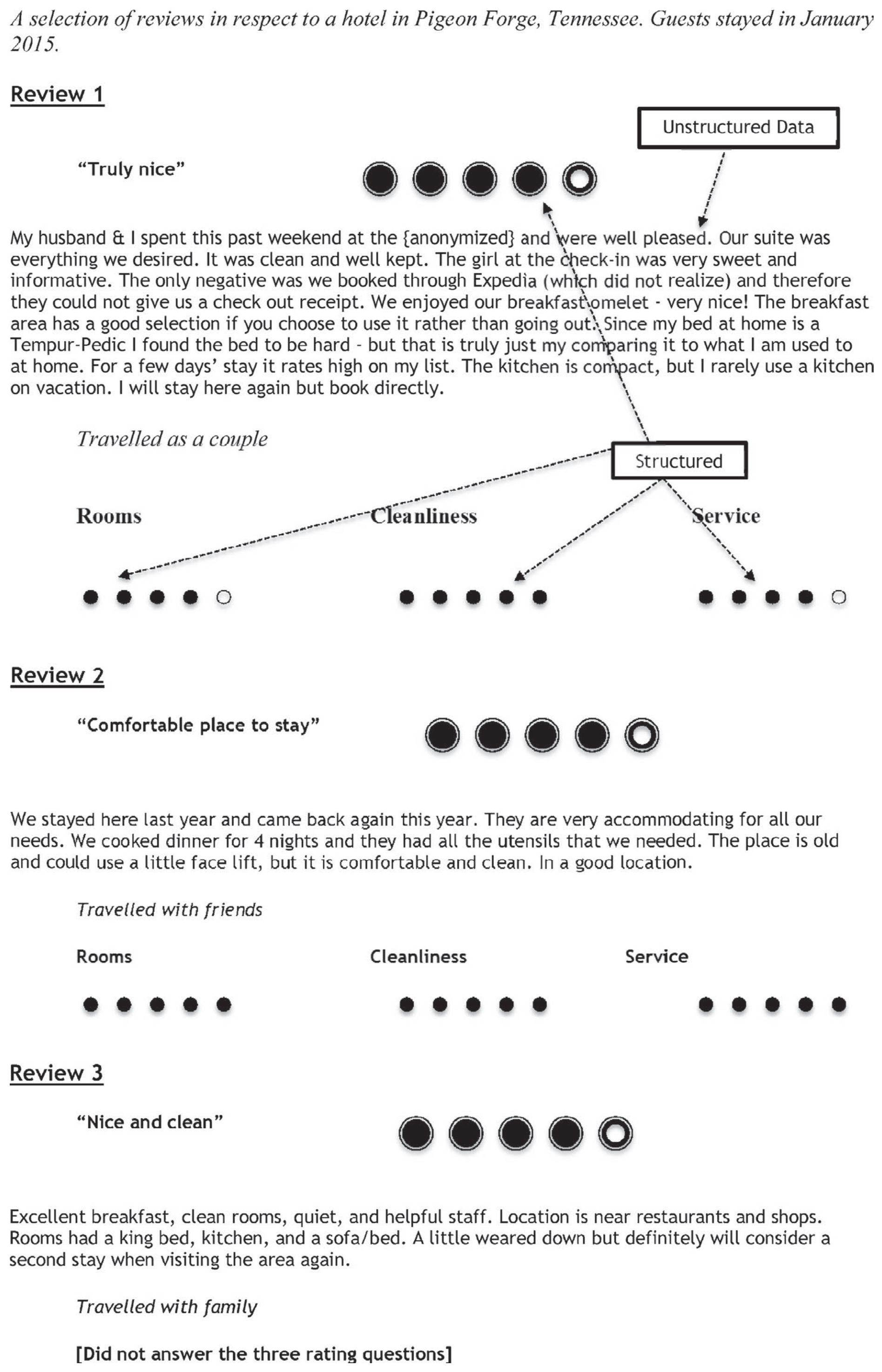
The strategic benefits of accessing large amounts of feedback are easy to imagine. For example, firms launching new products now have a significant opportunity to learn from the feedback of early adopters. The firm might consider a soft launch: a limited rollout with little fanfare. The product can then be tweaked upon gaining initial reactions. Online brand communities have great potential to tell us how selected groups of consumers see the brand. Marketers involved in the field of Online to Offline (O2O) sales also have a growing opportunity to trace the relationship between offline sales and online postings (e.g., tweets) about the product. Indeed, online postings may prove to be a leading indicator of sales, meaning consumer comments can assist firms well beyond the marketing function. Earlier and more accurate sales predictions will help logistics, production scheduling, and financial planning. The value of improved demand forecasting should be especially significant in industries with long lead times and/or where consumer tastes are hard to predict, such as fashion retailing and vacation planning.

1.2. A taxonomy of big data

Big data can be divided into two types: structured and unstructured data. *Structured data* comes in a defined form and offers clear answers. In user-generated content, structured data typically includes things such as ratings (e.g., 1—5 stars), questions with binary answers (e.g., "I would recommend Yes/No"), or questions with a limited range of responses (e.g., "Do you think the firm should do more, do less, or is the level of service about right?"). This data is similar to what marketers have been employing for generations and remains useful. It is generally relatively easy to extract the message from such data; for example, an average rating of 8.5 on an increasing scale from zero to ten is better than a rating of 7.5 in response to "How happy are you with our service?" This makes structured data especially useful when setting and monitoring performance against targets. It is easier to rally a team behind a simple, easy-to-communicate metric (Reichheld & Markey, 2011).

Unfortunately, structured data also suffers from many weaknesses. Look at the feedback on the hotel in Figure 1: some questions simply go unanswered. Users don't like being forced into choices, and instead ignore the question or pick randomly when unsure how to respond. This lessens the data quality,

Figure 1. **Examples of structured and unstructured data**



A selection of reviews in respect to a hotel in Pigeon Forge, Tennessee. Guests stayed in January 2015.

Review 1

"Truly nice" ●●●●○

My husband & I spent this past weekend at the {anonymized} and were well pleased. Our suite was everything we desired. It was clean and well kept. The girl at the check-in was very sweet and informative. The only negative was we booked through Expedia (which did not realize) and therefore they could not give us a check out receipt. We enjoyed our breakfast omelet - very nice! The breakfast area has a good selection if you choose to use it rather than going out. Since my bed at home is a Tempur-Pedic I found the bed to be hard - but that is truly just my comparing it to what I am used to at home. For a few days' stay it rates high on my list. The kitchen is compact, but I rarely use a kitchen on vacation. I will stay here again but book directly.

Travelled as a couple

Rooms	Cleanliness	Service
●●●●○	●●●●●	●●●●○

Review 2

"Comfortable place to stay" ●●●●○

We stayed here last year and came back again this year. They are very accommodating for all our needs. We cooked dinner for 4 nights and they had all the utensils that we needed. The place is old and could use a little face lift, but it is comfortable and clean. In a good location.

Travelled with friends

Rooms	Cleanliness	Service
●●●●●	●●●●●	●●●●●

Review 3

"Nice and clean" ●●●●○

Excellent breakfast, clean rooms, quiet, and helpful staff. Location is near restaurants and shops. Rooms had a king bed, kitchen, and a sofa/bed. A little weared down but definitely will consider a second stay when visiting the area again.

Travelled with family

[Did not answer the three rating questions] and thereby the insights that can be extracted from it. Often, customers have a more complex and nuanced view than is captured by the options in the structured data. For example, imagine that you have bought a product online. You love the product but delivery problems meant you were greatly inconvenienced: you wasted a morning waiting for the delivery to arrive and missed an enjoyable lunch date. Should you rate this product as five stars because when it finally arrived the product was wonderful? Or should you rate the product much lower because of the delivery hassle? Interpreting responses is challenging because different customers will make different choices. By summarizing complex thoughts in a single number, the reasons that informed the choice are lost. Look again at the reviews in Figure 1. All the reviewers gave four out of five ratings, but the discussions demonstrate that these customers have unique opinions. The rating alone won't help the manager turn this good hotel into a great hotel.

In contrast to structured data, *unstructured data* comes in a form that is amorphous and which must be treated in order to be usable. For example, reviews written in freeform English represent unstructured data. The aforementioned individual who loved the product but experienced delivery headaches could describe via customer review the brilliance of the product while also concurrently highlighting the delivery problems he/she experienced. Such unstructured data is usually rich and exciting, but can be extremely hard to use. Managers trying to estimate what consumers think from skimming these reviews will typically form heavily biased estimates. And when the volume of reviews is very high, reading all the reviews can present significant practical challenges.

Yet there is great information to be gleaned with the help of an appropriate approach. The reviews in Figure 1 highlight the strengths of the hotel, such as its location and helpful staff. No marketer had to prompt the customers with questions such as: ''What did you think of the location?'' With freeform reviews customers share their thoughts on what is important to them, not what market researchers predict will be important; this will often highlight areas of potential improvement. One reviewer suggests the beds aren't too comfortable. Two reviewers note that the décor looks a little old, but consider how they express this invaluable information: one suggests that the hotel could ''do with a face lift'' and another describes the hotel as ''weared down.'' It is clear what the reviewers mean, but the terminology used is non-standard English. This makes it hard for many traditional text-mining programs to uncover the message, especially those that rely on dictionaries of words selected in advance. In such cases market researchers need a technique that is flexible enough to group similar ideas together even when consumers use a variety of terms for the same idea.

One positive feature of unstructured data is that not only can analysts easily access it from their own company websites, but also it is often almost as easy to access data on sites run by third parties. Without the right methods, however, analyzing unstructured data seems akin to cleaning the Augean Stables: those attempting it will get messy and receive little thanks. One problem is that even among customers inclined to recommend a great product, the recommendations take substantially different forms. Some will applaud loudly, noting in exquisite detail how the product delivered a wonderful outcome. Other customers will less ostentatiously recommend with several well-chosen words. Other customers will post a laconic ''good job,'' metaphorically patting the firm on the back. To understand what customers' comments mean, we must transform the huge amount of text, all in different writing styles, into insights. We need a way of extracting the message from the mess of big data.

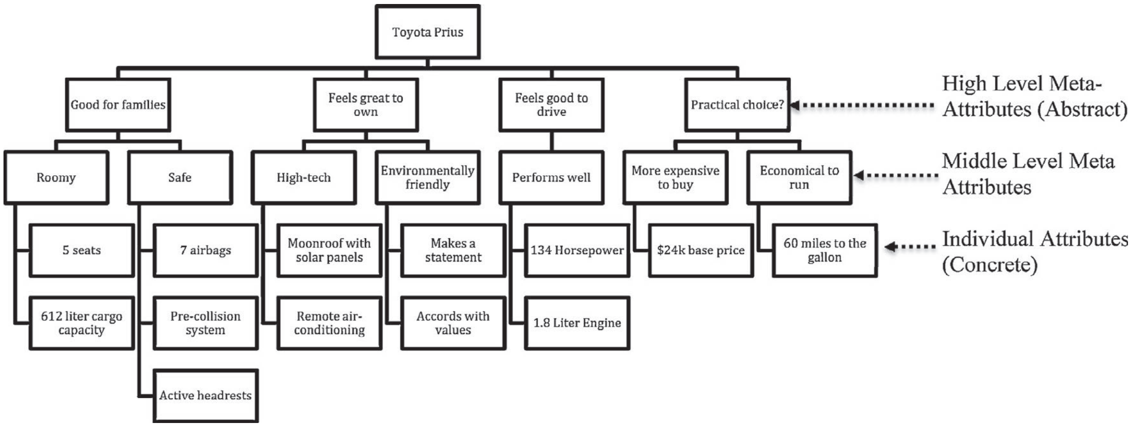
The challenges of unstructured data mean many managers don't know how to squeeze the information out of it. This article recommends a publicly available, non-proprietary way of doing so. Latent Dirichlet Allocation (LDA) can take a massive amount of text and extract the messages. While there is no simple, off-the-shelf application to run LDA, any firm employing analysts with statistical training should be able to implement the approach we recommend.

2. Uncovering the unseen attributes that matter to customers

The challenge in understanding the message from much user-generated content arises from the nature of unstructured data. For example, when reviewing a smartphone model, a customer's comments might meander from the size of the phone's screen to the weight of the device, move on to cover the availability of apps, digress to the speed of processing, and finally return to focus on the screen but describe its resolution rather than its size. Furthermore, many comments aren't interpretable out of context. For example, when a consumer mentions the weight of a portable tech device, this can be a positive or negative depending upon the words that accompany the reference to weight.

The problem we address here is that even when consumers value the same underlying feature,

Figure 2. Hierarchy of attributes



called an *unseen attribute*, they will often describe the attribute in different terms. For example, laptop speed may be a relatively simple attribute to consider (i.e., all customers understand what this means and prefer fast over slow), yet the diversity of comments remains intimidating. If we extract only the positive comments within this group, there will still be a great variety of terms used. Speed matters to Tony, a restaurant manager, but he doesn't give any thought to how this speed is achieved. He has been impressed by the speed of his new laptop, however, and leaves a recommendation to that effect. Tony's friend, Arnold, also wants a fast computer but he has an advanced layperson's knowledge of computers. Arnold believes it is the brand of CPU that drives the superior performance and happily comments that the CPU's brand makes this a great laptop for busy office workers like himself. Finally, Stephanie, Tony's tech-savvy wife, crunches data for a living and the CPU also matters to her. When leaving a review, she is careful to explain the precise specifications of the laptop, including the processor's family description. Although Stephanie is impressed by how fast her laptop is, she doesn't even mention speed but rather emphasizes the machine's impressive latency times. All three customers loved the laptop's speed but expressed this in very different ways. The review comments stretch from abstract generalities proffered by a lay person (Tony) to very concrete details provided by an expert (Stephanie).

The most concrete description possible entails listing technical specifications of the product. When reviewing an HDTV, experts will tell you how many lines of horizontal resolution there are (e.g., 1080p) and the number of frames per second (e.g., 60). Slightly less expert customers will typically discuss the same unseen attribute, picture quality, but describe it more abstractly, perhaps mentioning that the TV is "Full HD." The least expert customers will tend to be the least concrete, speaking of the unseen attribute (i.e., picture quality) at the highest level of abstraction, stating that the TV has a "great picture." In this scenario, there is a hierarchy (from concrete to abstract) with one high-level unseen attribute (picture quality) encompassing a multitude of more concrete comments. Figure 2 provides an illustration of the hierarchy of attributes that a consumer might hold for a Toyota Prius. Note that even when the consumer mentions safety, they may be focusing on the abstract idea of being safe or the concrete individual attributes related to safety, such as the pre-collision system.

Any method that seeks to understand the message must extract the common themes between the comments and so group together unseen attributes expressed at various levels of abstraction. How then can disparate words be grouped into unseen attributes? One effective way of grouping these comments together is via Latent Dirichlet Allocation (Blei, 2012).

3. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a procedure that extracts themes from a dataset (Griffiths & Steyvers, 2004). For user-generated content, these themes are the unseen attributes that each customer discusses. LDA has similarities to the principal components methods performed on customer surveys in that it extracts common underlying attributes.[1] Unlike many approaches, LDA copes well with the sort of messy data that managers have when

[1] These are also known as latent attributes, hence the 'L' in the LDA acronym.

they analyze user-generated content. The method's advantage is that it works with massive amounts of unstructured data where each customer comment does not address most attributes. For example, while only one in ten consumers might mention a smartphone's weight, LDA can still extract and analyze those views about weight that are expressed. Survey methods cannot ask about the wide range of things that LDA can pick up from analyzing user-generated content. While LDA is, by necessity, somewhat technical, its implementation is relatively easy for statistically trained analysts. The data necessary to conduct the analysis can be gained from sites such as Facebook and Twitter using the application programming interface (API) that these companies provide. If the company does not have the in-house expertise to secure the user-generated content data from public sites, the data can easily—and relatively cheaply—be purchased from third parties.

When analyzing the messy data, we start with the assumption that there is a hidden structure to the data—a set of unseen attributes—which underpins the comments. For example, we may assume that ten unseen attributes underlie the reviews of a certain product. The precise number can be determined by examining which model best fits the data. We use a method of minimizing perplexity, which creates a model on a subset of the data and tests how well the model describes the rest of the data. The idea is to analyze enough unseen attributes to provide a good prediction of consumer attitudes but not overfit the data to random occurrences in your data. While technically it is no problem to search for a very large number of attributes, allowing an excessive number of attributes to be considered means that messages consumers are sending you may get lost amongst the vast number of attributes considered.

The model analyzes all the comments and describes each comment as representing a combination of themes: the unseen attributes. For simplicity of explanation let us consider a product where we find that we can describe the data with only three attributes; of course, real-life data will typically have many more attributes than this. A customer review of a graphic novel might use the following unseen attributes: (1) storyline, (2) characters, and (3) graphics. Each review is unique and each review can be said to use a different proportion of each attribute. My review might be described as 60% about the storyline, 20% about the characters, and 20% about the graphics. Your review might split 30% about the storyline, 40% about the characters, and 30% about the graphics. We assume what the consumer proportionately decides to talk about in the review reveals what is important to each individual consumer. By aggregating across the entire dataset we uncover what matters to consumers as a group and here can tell the weight that graphic novel readers put upon storyline versus characters versus graphics.

LDA is a type of a wider class of statistical models called *topic models*. These can be viewed as attempting to reverse engineer the thinking of the person who wrote the text; from the words used, we try to infer the attributes the person was thinking about. Essentially, we ask: What is the most likely combination of attributes to have been in the consumer's mind for them to have written the review that we observe?

To understand the intuition behind LDA, imagine that the user-generated content being analyzed was generated randomly by drawing words from a given hidden structure. The structure can be represented as buckets that each contain words relevant to an unseen attribute. For graphic novels, 'plot' or 'narrative' might be words that are in the bucket that we label storyline, whereas the words 'drawing' or 'illustration' might be associated with the bucket that we label graphics. The hidden structure that LDA uncovers specifies the proportions of the unseen attributes; for example, how often each consumer draws from each bucket. We assume that the more a consumer uses words related to an attribute, the more important the attribute is to the consumer. This reasonable assumption can, however, be somewhat misleading if the consumers are for some reason focused on less important attributes; for example, if the consumer is prompted to write about certain attributes by the review website.

How does LDA uncover the hidden structure? The model tests all the possible hidden structures that could exist given the number of unseen attributes the analyst previously decided fit the data. The model then determines the likelihood that each hidden structure could have created the data observed using the Dirichlet distribution, a distribution often used to represent estimated probabilities. The hidden structure found to have the highest probability of having created the data is then identified as being the underlying hidden structure (Wang, Bendle, Mai, & Cotte, 2015).[2]

LDA uses at least three critical basic ideas. First, words contained in each piece of text are generated from a list related to each unseen attribute. For reviews of the Prius, 'seats,' 'cargo,' and 'capacity'

[2] For details on the technical procedure used, see the online appendix to Wang et al.'s (2015) work.

might be words in the list that could be used in reference to the attribute roomy. Second, each unseen attribute has a probability distribution over a fixed word vocabulary, so those consumers thinking about the roominess of a car are likely to mention cargo at a given probability. Third, the unseen attributes are shared by all of the pieces of content, but proportions differ. This means that all consumers could potentially think about the same things: we all could discuss the roominess of the Prius, but each individual may or may not do so. Technically, LDA goes through a number of stages in its estimation:

- It estimates the length of the piece of user-generated content using a distribution of the lengths of content. This is important because user-generated content can vary widely in length. For example, some online reviews use one word (e.g., excellent) while others are much more verbose. Clearly, more attributes can be discussed in a longer piece.
- LDA samples the proportions of the attributes discussed using possible distributions of attributes. This is important because the consumers often vary considerably in what they care about.
- For each of the words in the piece of user-generated content, LDA samples from the attributes and the list of words associated with that attribute.

When LDA was developed it only looked at attributes (e.g., storyline) but did not consider if the attributes were discussed positively or negatively. The method can, however, be adapted to consider the valence of comments (i.e., positivity/negativity). For example, references to the weight of a smart phone may be positive or negative depending upon context. Tirunillai and Tellis (2014) show how to uncover the difference between positive and negative references to the unseen attributes. The analyst ‘seeds’ the LDA model by entering a number of clearly valenced words (e.g., good, bad). The model then assigns unseeded words a valence (positivity/negativity) depending upon the regularity with which they appear beside the seeded words. Words commonly occurring near references to ‘good’ are assumed to describe positive features of the product. A word that might be ambiguous without its context, such as ‘lightweight,’ is detected by the model to be positive (in the case of a smartphone) or negative (for patio furniture) depending upon the words with which it regularly co-occurs in the dataset.

3.1. The benefits of LDA

The Latent Dirichlet Allocation (LDA) model we recommend here is not the first for textual analysis. Such models tend to come with a variety of exciting names and enigmatic acronyms such as Naïve Bayes Classification and Latent Semantic Analysis (LSA). In general, the strength of the LDA process is that it requires fewer ad-hoc assumptions than many other approaches. It builds upon a rigorous foundation of statistical inference and extends the ideas of probabilistic Latent Semantic Analysis (Blei & Lafferty, 2009). LDA is unsupervised: the analyst lets the statistical properties determine the unseen attributes. Unsupervised processes give flexibility but make the computation lengthier than in supervised processes. A strength related to its flexibility is that LDA does not need to employ a dictionary or thesaurus (Hofmann, 2001). This is especially advantageous when, as we saw in the reviews of the hotel (Figure 1), people misspell words or revert to colloquial language. LDA isn’t reliant on anyone maintaining an ever-growing list of words that consumers may choose to use, or misuse, online in increasingly unpredictable ways.

A particular quality worth noting is that LDA provides ‘soft’ classification for each piece of user-generated content. We do not simply classify a review into a single category: *This customer cares about storyline*. Instead, LDA provides a more nuanced view: *This review focuses on the storyline but also highlights that the characters were incredible despite the graphics being disappointing*. LDA sees each piece of unstructured data as a mixture of different topics and so better captures the complex views consumers hold. The consumer writing the review isn’t just bucketed as a promoter or a detractor. We can delve further into why the consumer is happy—or not so happy—with the product to better understand what underlies the consumer’s assessment. A consumer will often like some things, but not like other things. With LDA we can hope to understand this from what is said in public forums.

3.2. Challenges for LDA

As with any technique, LDA is not perfect: assumptions have to be made to analyze big data. One key assumption is that reviews are a ‘bag of words.’ Essentially, this means that the order of words does not matter. Each piece of user-generated content is represented simply as a list of word counts; for example, in reviews of a textbook, ‘boring’ was mentioned 13 times while ‘witty’ was only mentioned once. Such a list neglects the order of the words so LDA cannot assess any difference between a review that mentions, for instance, screen size at either the beginning or the end of the review. In many situations this will not matter. That said, LDA throws away information about word order to extract the messages. If, for example, a review website asks consumers to lead with the information they find most critical to their enjoyment, then word order matters and LDA is probably not the right technique to use. A similar issue occurs when consumer perceptions change over time. LDA will give an average perception across the entire period that may not represent the state of consumer opinion at the end of the period. As such, analysts should try to ensure that all the pieces of content analyzed remain relevant.

Unsupervised, as compared to supervised, methods rely less on the specialist market knowledge an analyst possesses. This reduces the chance of analyst bias influencing the results. It also means that the model ignores potentially valuable information held by the analyst. LDA extracts themes, but the analyst names the themes, so despite being unsupervised LDA cannot totally remove the chance of analyst bias influencing the presentation of results.

LDA has two further technical limitations that are worth considering. First, no topic correlation is allowed. In our graphic novel example, storyline is assumed to be distinct from characters, as is graphics. It is possible that, for example, storyline and characters share more similarities with each other than they do with graphics. LDA will not spot such similarities. Finally, LDA is a practical development from computer science. It lacks a clear theoretical justification for some of the decisions that must be made. For example, there is no theory that can be applied that tells an analyst how many themes to search for given the number of documents analyzed. While some choices will be easier to defend than others, there remains considerable subjectivity in the analyst's choices.

Accessing data from outside servers also requires skills to scrap the data. Numerous third parties will do this for a reasonable price. That said, the cost may limit some firms' ability to procure regular data in a timely manner, and regular data extraction is critical to gaining insights from LDA when there are major changes in the thinking of consumers over time.

Perhaps the most crucial barrier to implementation is that the analyst must understand the technique. While LDA is relatively easy—indeed, we believe anyone with a postgraduate knowledge of statistics will be able to implement LDA after referencing the work cited in this article—it still requires some statistical skill. Large corporations should have little difficulty finding people with the requisite skills, but this may be more challenging for smaller organizations.

As with any technique to analyze user-generated content, LDA can only uncover what consumers choose to share. The good news is that consumers are becoming increasingly open to sharing their thoughts. It still remains a possibility, however, that those who share their thoughts differ in some way from those who do not. Given the limitations that we have highlighted, we wish to be clear that we are not advocating LDA as a firm's only form of market research. For example, a firm may want to validate the findings of LDA by conducting research with a representative sample of consumers. Such a test could look for any bias caused by the consumers who generated the content being in some way atypical of the general population.

Despite the challenges facing LDA, we suggest that it can be an important part of many firms' methods of monitoring consumers' opinions. It can be performed in-house, inexpensively, regularly, and in a timely fashion on data that is readily available.

4. What managers can learn from Latent Dirichlet Allocation

4.1. What matters to customers in your industry

LDA reveals the hidden structure of unseen attributes behind comments. This allows greater insight into what consumers are thinking. Market researchers no longer have to somehow determine which attributes consumers care about before surveying the consumers using questions related to those attributes. The firm can uncover what matters to consumers from their own unfiltered words. Given that even before entering an industry a firm can accurately estimate what consumers in the industry care about, this should make the playing field more level for new entrants.

LDA makes progress toward solving a problem with traditional surveys, namely that sometimes consumers can't express what they think when put on the spot during an interview. While user-generated content typically won't discuss potential innovations that the consumer isn't aware could be offered, an analyst can see what the consumers care about in the current market offerings at various levels of abstraction. This will allow firms to better predict how consumers will respond to any novel features that the firm is thinking of introducing to the market.

4.2. What customers are saying about your products

LDA allows us to categorize consumer reactions. Armed with this understanding, we can know in which areas a firm is satisfying its customers. The exercise thus tells management which areas of the customer experience require attention. LDA's ability to deal with unstructured comments gives control of the agenda to consumers, leading to benefits for both consumers and firms. The consumers get to share what they are interested in without having to answer structured questions, which we all sometimes find challenging or tedious. The firms get to understand consumers' concerns more fully, which allows any costly improvements undertaken to be more efficiently targeted on things that matter to the consumers.

4.3. The market structure of your industry

LDA is a flexible procedure that, when mastered, can be applied to analyze not just your own firm's position in the market but also the position(s) of your competition. An analyst can access data on any number of competitors from publicly available user-generated content, such as online reviews of competitors' products. LDA is then able to identify the concerns that matter most to the customers of each competitor in the market. For example, customers of Waitrose, a UK grocer, might be seen to care more about local sourcing than do customers of a competitor. Understanding the importance that consumers assign to each unseen attribute allows a marketer to place products/brands/firms on a multi-dimensional map showing which unseen attributes are most associated with each competitor. This allows firms to assess who their closest competitors are where it counts most: in consumers' minds.

4.4. The weaknesses of your competitors

Until rather recently, consumers' opinions regarding the strengths and weaknesses of a product or service could only be discovered via personal survey. Now, however, market research can be conducted by downloading reviews from websites like Amazon.com and tripadvisor.com, and analyzing what consumers are saying about rival firms. This not only enables an understanding of the differences between what customers of various firms care about, but also enables the analyst to dig deeper. We can uncover which firms are simply not satisfying their customers based on what the customers say. For example, from comments posted on Twitter, what is Comcast doing badly? What is the company doing well?

As high-quality, real-time data on performance becomes more widely available, each firm has every incentive to improve. The profusion of online reviews and other user-generated content makes it increasingly harder for firms to hide failures. Consumers benefit whenever it becomes more burdensome for firms to try to hide their areas of weaknesses than it is to simply improve those areas instead. Furthermore, consumers are more likely to reward good—and punish poor—performance as a firm's good and bad points become more widely known. This will help those firms that offer absolute value (Simonson & Rosen, 2014). We hope that LDA can increase competition on things that matter to consumers by allowing firms to understand more quickly where they need to improve.

LDA is not only a tool to better understand how people see your product, but also a critical tool in competitive strategy. It allows managers to uncover how successful a competitor is at satisfying its customers. Critically, understanding how a competitor is failing its customers can expose areas in which your rival's customers may be tempted to defect.

4.5. That unstructured data may be intimidating, but it can be tamed

Big data is a popular topic. Many managers will have given some thought to how it impacts them, but the sheer messiness of big data can be intimidating. Presented with over 10,000 reviews of your product, there is a temptation to skim a few and hope you get the gist. Sadly, however, the human brain isn't great at summarizing large amounts of data without bias. Such crude methods of analysis are liable to miss many important stories, so we suggest that you should aim to tame the big data. This is what using Latent Dirichlet Allocation allows you to do: extract the message from messy, unstructured, big data.

While most managers probably realize that powerful messages lie within the mess of big data, many don't know that these messages can be extracted relatively easily using the right techniques. While we are recommending a specific technique that can be implemented today by firms wishing to do so, the central message of our article isn't that all should adopt LDA immediately. Rather, we highlight that LDA—or a similar approach—can now be implemented. Managers should seriously consider this; remember, your competitors may be doing so already!

Techniques are advancing every day, so other methods of analysis will surely overtake LDA at some point. We predict that successful firms will be those that can use the best techniques available to extract the message from big data. The exciting point is that each technical advance makes seams of data that were previously uneconomical to mine available to be exploited. While none of the available big data techniques are perfect, they can help bring firms closer to the consumer. Our goal is to emphasize that we can understand the messages being sent by consumers if we use the right tools to listen.

5. Conclusion

Firms have easy access to data regarding the performance of their products, what consumers really care about, and the strengths and weaknesses of competitors. Consumers are not shy about sharing their thoughts on any number of topics via public forums. This user-generated content contains incredible potential, but many firms don't know how to properly tap it. We suggest that firms consider Latent Dirichlet Allocation, a non-proprietary technique that can be applied by anyone with advanced statistical training. This allows analysts to extract what consumers are thinking about from user-generated content. This technique even allows a manager to understand which attributes consumers see as positives or negatives of his/her product and competitors' products. Such analysis can inform the firm's strategy to better serve consumers. With the right tools, the message can be extracted from the mess of big data.

References

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77—84.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71—94). Boca Raton, FL: CRC Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228—5235.

Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning, 42*(1/2), 177—196.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Special report: Analytics and the new path to value. *MIT Sloan Management Review, 52*(2), 22—32.

Reichheld, F., & Markey, R. (2011). *The ultimate question 2.0: How net promoter companies thrive in a customer-driven world*. Cambridge, MA: Harvard Business Press.

Russell, C., & Bennett, N. (2015). Big data and talent management: Using hard data to make the soft stuff easy. *Business Horizons, 58*(3), 237—242.

Simonson, I., & Rosen, E. (2014). *Absolute value: What really influences customers in the age of (nearly) perfect information*. New York: Harper Business.

Thelen, S., Mottner, S., & Berman, B. (2004). Data mining: On the trail to marketing gold. *Business Horizons, 47*(6), 25—32.

Tirunillai, S., & Tellis, G. (2014). Mining marketing meaning from chatter: Strategic brand analysis of big data using Latent Dirichlet Allocation. *Journal of Marketing Research, 51*(4), 464—479.

Wang, X., Bendle, N. T., Mai, F., & Cotte, J. (2015). The journal of consumer research at forty: A historical analysis. *Journal of Consumer Research, 42*(1), 5—18.