Article

International Journal of Market Research

Trend spotting: Using text analysis to model market dynamics

International Journal of Market Research
2018, Vol. 60(4) 408–418

DOI: 10.1177/1470785318758558
journals.sagepub.com/home/mre

Jameson Watts
Willamette University, USA

Abstract

As marketers, we are confronted by an increasingly complex world—one characterized by markets that emerge and evolve at an unprecedented rate. This has led to an increased need for methods that can help researchers translate this complexity and dynamism into actionable intelligence. In the current work, I introduce one such method, and show how it can be applied to the emergence and evolution of the biotechnology industry during the 1990s and early 2000s. I conclude with a discussion of how this method can be applied more broadly to areas like content marketing, trend spotting, and the interface between qualitative and quantitative market research.

Keywords

big data, information theory, natural language processing, sociocognitive structures

Introduction

> Grammars in both cooking and engineering exist not just as rules but as a set of unspoken practices taken for granted. (Arthur, 2009, p. 77)

Over the past decade, we have witnessed a growing interest in the study of markets and their dynamics (Dolbec & Fischer, 2015; Humphreys, 2010; Peñaloza & Venkatesh, 2006). This interest is fueled by the realization that market emergence and change are now occurring on an unprecedented time scale. Technologies like the Internet and cell phones have greatly increased the scope and rate of exchange, which in turn have increased the variety and variability of market conditions (Mele, Pels, & Storbacka, 2015). Addressing this dynamism is paramount for researchers interested in understanding the forces that shape the modern marketplace (Reibstein, Day, & Wind, 2009) and for practitioners interested in methods that can transform this complexity into actionable intelligence (Rolland & Parmentier, 2013; Schmidt, 2010).

In response to this need, a variety of scholars have proposed theories and methods suited to this growing dynamism. Noteworthy among these efforts is a group of researchers focused on the

Corresponding author:
Jameson Watts, Atkinson Graduate School of Management, Willamette University, 900 State Street, Salem, OR 97301, USA.
Email: jwatts@willamette.edu

sociocognitive properties of market constitution (see, for example, Dolbec & Fischer, 2015; Humphreys, 2010; Peñaloza & Venkatesh, 2006; Vinhas et al., 2010). As markets emerge and change, new ideas and behaviors also emerge, spread, and eventually form the basis for a shared understanding of what a particular market is, and how it operates (Humphreys, 2010). When this type of understanding (of a product category, behavior, norm, etc.) is broadly shared, it can be called a sociocognitive structure: a cognitive representation of reality that is shared by many social actors.

Consider, for instance, the introduction of the minivan. When the market first emerged, very few individuals understood what a minivan was, or how it might be incorporated into their lives (Rosa, Porac, Runser-Spanjol, & Saxon, 1999). However, over time, more and more market actors developed similar ideas about what a minivan was supposed to do, what it cost, and how you were supposed to feel when you own one. When viewing the world through a sociocognitive lens, a market emerges when actors "come to certain shared understandings of what is being exchanged and why" (Humphreys, 2010, p. 2).

Within this stream of research, most empirical studies analyze the emergence of markets (and their concomitant sociocognitive structures) by looking at the exchange of ideas that led to the shared understanding in the first place. More often that not, these exchanges involve discourse (Lawes, 2002). Language, it is argued, is the device through which notions of appropriateness are contested, negotiated, and renegotiated (Hardy, 2011; Powell & Colyvas, 2008)—a process that eventually leads to the establishment of field-wide knowledge and shared understanding of the boundaries and definition of relevant technological dimensions (Santos & Eisenhardt, 2009), the meaning of new product categories (Rosa et al., 1999), and the validity of new practices (Lounsbury & Crumley, 2007; Tripsas, 2009). Observations of language at different points in time should therefore provide insight about the level of shared understanding in the marketplace and, hence, the state of market development.

As a methodological device, the use of language in the measurement of shared understanding shows promise. However, current implementations lack a coherent theoretical rationale. For example, sociocognitive structures are often defined ex post—operationalizations are based on the present incarnation of a legitimized category and then traced to some earlier period. Petkova, Wadhwa, Yao, and Jain (2014) typify this practice when they examine investing behavior in the clean energy sector by counting historical references to the terms "clean energy," "green energy," and "alternative energy" in media articles. Yet, the meaning and significance of these terms is only credible after a market has matured. In practice, categories of behavior emerge organically (Goldberg, 2012) and may cycle through several incarnations (and labels) before they settle on the perspicuously demarcated boundaries that appear to the present observer.

Furthermore, increases in shared understanding may in fact decrease the number of times a keyword is used in communication. This would occur if, for instance, a behavior is understated by virtue of its overwhelming taken-for-grantedness. For example, patrons of North American restaurants are unlikely to mention a chair as part of their review of a dining establishment. Yet we would be remiss to discount the importance of a chair to their experience—a lack of suitable options for sitting would almost certainly be deemed inappropriate and the behavior viewed as inconsistent with the broadly shared understanding of what a North American restaurant is.

How, then, do we use language to detect the emergence and change of these sociocognitive structures in an unbiased manner? First, the approach needs to remain keyword-agnostic so as to skirt issues of researcher bias. Second, it needs to account for the fact that a sustained low occurrence of certain words or phrases (like "chair") is also meaningful. To accomplish these objectives, I propose a method for the measurement of shared understanding that stems from changes in the way language is used to describe a given topic. This can be accomplished without imposing ex post

classifications or inflating the importance of high-frequency terms by tracking changes in the distribution of descriptive words across time.

The validity of this approach rests on the argument that consistency in language use is a reasonable proxy for shared understanding. To illustrate, imagine two scholars from the same discipline discussing a recent research article. For well-known concepts, often a single word or phrase is adequate to convey the intended meaning. More importantly, participants will use the same word or phrase each time the concept arises such that a series of similar conversations will exhibit a high degree of language consistency.

Contrast this with a situation in which a scholar who specializes in one academic area is explaining a research topic to a scholar from another area. In this scenario, each participant uses a variety of words and phrases to both ensure and demonstrate understanding of the content. Repeated observations of such discussions would show a gradual increase in language consistency as the words and phrases used to describe the concept are learned and their meaning agreed upon. As Arthur (2009) in the opening quote recognizes, once the context and meaning of terms or phrases is taken for granted, communication is less burdened by explanation.

Other researchers have made analogous connections between language use and shared understanding. Most notably, Suchman (1995) describes the development of shared understanding as a process accompanied by discourse that evolves from active evaluation of substantive claims to passive support of the dominant justification. For Green, Li, and Nohria (2009) and Harmon, Green, and Goodnight (2015), this transformation is manifest in the simplification of sentence structure over time—a process of compression in which lengthy and varied discourse is supplanted by more compact text. When an argument is new, actors must actively engage in justification. Over time, however, a claim can be taken for granted to such an extent that "for things to be otherwise becomes literally unthinkable" (Zucker, 1983: 25).

In the subsequent sections, I describe my language-based model of shared understanding. Using textual data from 13 years of trade journal articles, I then construct a measure derived from this model—which I call "language consistency"—and track its evolution over the formative years of the biotechnology industry. Consistent with contemporary descriptions of market emergence (Hakala, Nummelin, & Kohtamäki, 2017; Harmon et al., 2015), I find that language consistency generally increases over time (i.e., as the industry matures). However, I also find that many of the drastic drops in language consistency correspond with important market events. I conclude with a discussion of how this method can be applied more broadly to address contemporary issues in market research like spotting trends, constructing promotions, and tracking market development.

Data and measures

> I would argue . . . that the concept "redundancy" is at least a partial synonym of "meaning." As I see it, if the receiver can guess at missing parts of the message, then those parts which are received must, in fact, carry a meaning which refers to the missing parts and is information about those parts. (Bateson, 1972, p. 420)

Linguistic context

A variety of recent studies have started to look at the generic characteristics of written language as a way to understand the structure of markets and the evolution of human behavior (Goldberg, 2012; Klingenstein, Hitchcock, & DeDeo, 2014; Murdock, Allen, & DeDeo, 2015; Tirunillai & Tellis, 2014). The viability of this type of research is driven by parallel developments in the availability of digital archives and methods appropriate for the analysis of "big data." While the standards and

practices used for large-scale textual analysis are still very much under development, I have largely followed the work of prior scholars in preparing my data and constructing my measures. The following paragraphs describe this process in detail.

Textual data are sourced from articles written for the Bioworld trade journal between January 1991 (when the journal began) and January 2004. By 2004, academics and practitioners largely viewed the biotechnology industry as fully "legitimized" and mature as evidenced by broad isomorphic behavior and declining interest in data describing its development (K. W. Koput, November 11th, 2016, private interview). However, during the period between 1991 and 2004, Bioworld constituted a primary source for industry-wide dissemination of information about the activities of firms and other stakeholders in the biotechnology industry (Powell, White, Koput, & Owen-Smith, 2005; Wolff, 2001). The journal published several articles each day (excluding weekends) on topics as wide-ranging as personnel changes, fundraising activity, inter-firm contracting, the Food and Drug Administration (FDA) approval process, and the latest scientific trends.

Frequency distributions

Between 1991 (when the journal first started) and the end of 2003, Bioworld published 20,999 articles. Following the work of Klingenstein et al. (2014), Tirunillai and Tellis (2014), and others, I performed the following steps on each article in order to prepare the data for further analysis:

1. Break apart each article into a list of lexical elements at the word level in a process typically referred to as tokenization (Jurafsky & Martin, 2000).[1]
2. Remove all punctuation and numbers from the list.
3. Change all elements to lower case.
4. Remove all English stop words like "and," "the," "what," and so on (Jurafsky & Martin, 2000) and words shorter than three letters.
5. Stem the remaining elements using the Porter (1980) stemmer so that words like "work" and "working" are treated as the same lexical element.

Following the preprocessing steps described above, the lists of lexical elements representing each article were pooled by month into 156 larger lists—one for each month from January 1991 through December 2003. From these lists, I then constructed 156 frequency distributions, with the lexical element in rank order on the x-axis and its count divided by the total number of lexical elements constituting the y-axis—that is, that element's probability given the total list of elements in that month.

Kullback–Leibler divergence

In computational linguistics, a process called topic modeling is gaining in popularity as way to extract meaning from large, unstructured textual databases (Blei, 2012). In its most basic form, the process involves the naive classification of lexical "features" into buckets based on criteria such as co-occurrence within the text. For instance, the terms "player" and "court" might co-occur in many articles about basketball and thus contribute highly to that topic (Blei & Lafferty, 2009).[2] Such techniques have been successfully implemented across the social sciences to understand research agendas as varied as the classification of scientific knowledge (Blei & Lafferty, 2007) and brand positions in a competitive market (Tirunillai & Tellis, 2014).

Nonetheless, the process under investigation in the current work is dynamic, and the most common methods for naive topic modeling produce classifications that are static across time—the order of documents is irrelevant to the production of the classificatory scheme. Even the dynamic topic

model recently proposed by Blei and Lafferty (2006) imposes some restrictive assumptions. For instance, the set of features upon which the dynamic model is estimated must be defined prior to execution of the algorithm (Blei, 2012). A typical procedure is to choose the total set of features as the union of the top X features by frequency from each time period. However, this allows the most prominent features that exist far into the future (i.e., vocabulary) to influence topics modeled in the distant past. Moreover, the insights one gains from such models are based on how the various features change in their influence of a persistent topic—a worthwhile endeavor, but limiting in the current context where topics are assumed to emerge, merge, and disappear over time (Goldberg, 2012).

An alternative approach—and the one pursued here—imposes a somewhat less burdensome restriction. Rather than define the set of features using the entire (time-invariant) corpus, I draw from a moving window around the period under consideration. Changes are captured as the difference between features at time t and the features defined by their average over the previous k periods. Thus, I can limit features to the intersection of the top X by frequency in periods $t - k$ through period t. Moreover, the k-period moving average incorporates innovations gradually so that new shocks are defined against the relevant past—the period of time most likely to exist in an actor's recent memory—rather than the field's entire history.

The features I consider for my model are word-stems. They are mapped to a probability distribution based on the frequency with which they occur in a given time period. This is sometimes referred to as a "bag-of-words" model because it employs word-stems and their frequencies without consideration of their contextual ordering in sentences, paragraphs, and so on (Jurafsky & Martin, 2000). I compute the uniqueness of the distribution at period t by means of the Kullback–Leibler divergence (KLD) from the average of the prior k distributions (see, for example, Klingenstein et al., 2014). The KLD is denoted $D_{KL}\left(P \,\|\, Q\right)$ where P is assumed to represent the "true" distribution, which in this case is the distribution constructed from the prior k periods. Q is the distribution at time t. For discrete distributions (as I have here), the measure is defined as

$$D_{KL}\left(P_k \,\|\, Q_t\right) = \sum_i P_k\left(i\right) \ln \frac{P_k(i)}{Q_t(i)} \tag{1}$$

which describes the logarithmic difference between the probabilities P and Q. The consistency of this measure over time relies on the fact that the shape of the frequency distribution does not change appreciably with changes in the text—a fact grounded in Zipf's law (Adamic & Huberman, 2002; Zipf, 1932). Thus changes are largely based on a reordering of features within the distribution rather than a reconfiguration of its shape.

The KLD quantifies the amount of information lost when Q is used to predict P. As such, it captures the new information represented by the current set of word-stem frequencies. When the value is high, it describes a departure from the prior k distributions. However, when the value is low, it represents consistency with prior word-stem frequencies. Because the interest lies in similarity (rather than surprise), I can define a measure of language consistency over time as

$$LCON_t = -D_{KL}\left(P_k \,\|\, Q_t\right) \tag{2}$$

Analysis

For the analysis that follows, my measure of language consistency is based on probability distributions derived from the intersection of the top 1,000 word-stems in the current month (Q_t) and the prior three (P_k). Several other configurations were tested as part of my robustness checks and the results are qualitatively the same.

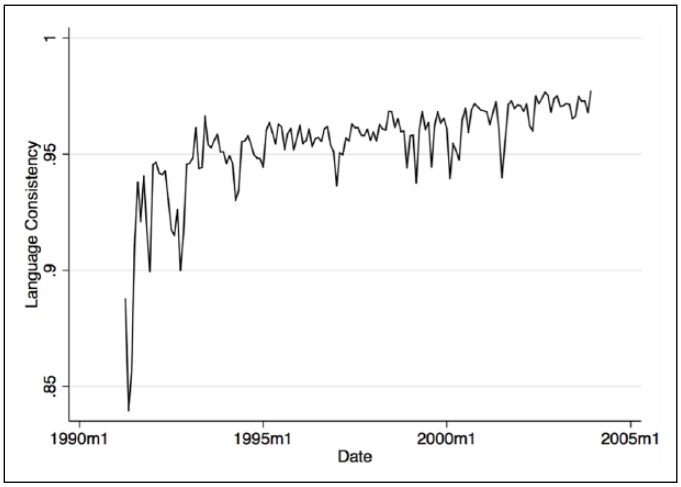

Figure 1. Monthly language consistency (1991–2004).

Several trends are immediately apparent. First, the amount of writing grew substantially throughout the 1990s before leveling off around year 2000. Second, the vocabulary used to describe the industry more than doubled over the same period with the sharpest increase occurring between 1992 and 1993. Third, the consistency in the way the language was used rose gradually over the same period. As argued in the prior sections, a rise in consistency suggests that the industry as a whole has undergone a gradual march toward the establishment of greater field-wide knowledge and shared understanding (Cattani, Ferriani, Negro, & Perretti, 2008).

Despite the gradual increase in language consistency year over year, there is significant variation at the monthly level. Figure 1 shows the language consistency measure plotted on a monthly basis from 1991 through 2003. Notably, there is significant volatility in the first couple of years, which may be due to institutional factors unrelated to the legitimation process. For instance, new reporters must be recruited and learn how to write about the industry. As mentioned above, the vocabulary more than doubled in these first few years. Further discussion is on the data presented after removing these first 2 years.

Figure 2 shows the language consistency measure for the years 1993 onwards in both levels (top) and first differences (bottom). However, there are several additional features worth noting. The upper graph shows several downward spikes, which represent strong deviations from the prior 3 months of language use. In several cases, this is followed by a rise back to relative consistency, suggesting that the change in language was incorporated and adopted. In other words, some industry shift took place, which then became a permanent feature of the lexicon.

One notable exception is the years 1999 through 2001 during which the consistency of the language fluctuates several times. This is most apparent when looking at the (lower) graph in first differences around this time. Clearly, the volatility of this measure is pronounced. This coincides with the general turmoil surrounding the "dot-com" bubble and the subsequent spillover into other industries. For instance, many biotechnology firms went out of business during this period for lack of financing, while others sought the safety of alliances with large pharmaceutical companies (Wolff, 2001).

Figure 3 highlights the most significant drops in consistency over the 11-year period (1993–2003) after which there appears a period of relative consistency in the language. The three

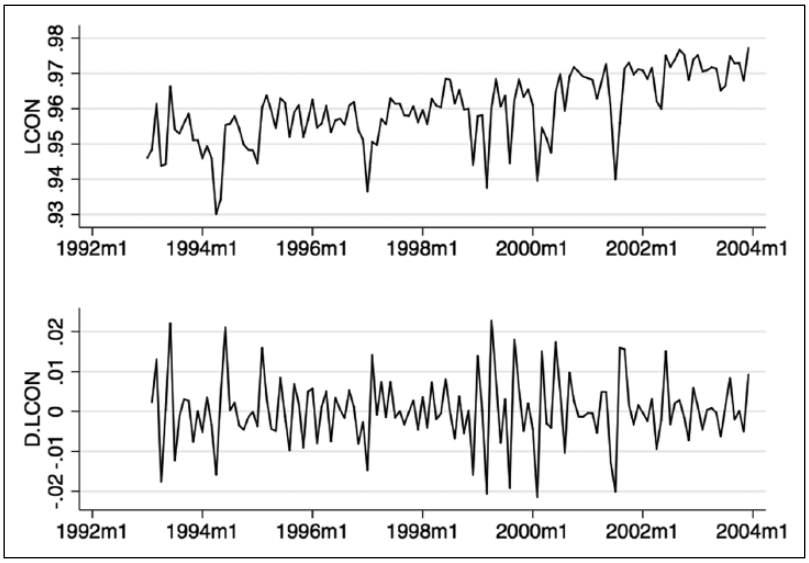

Figure 2. Monthly language consistency in levels and first differences (1993–2004).

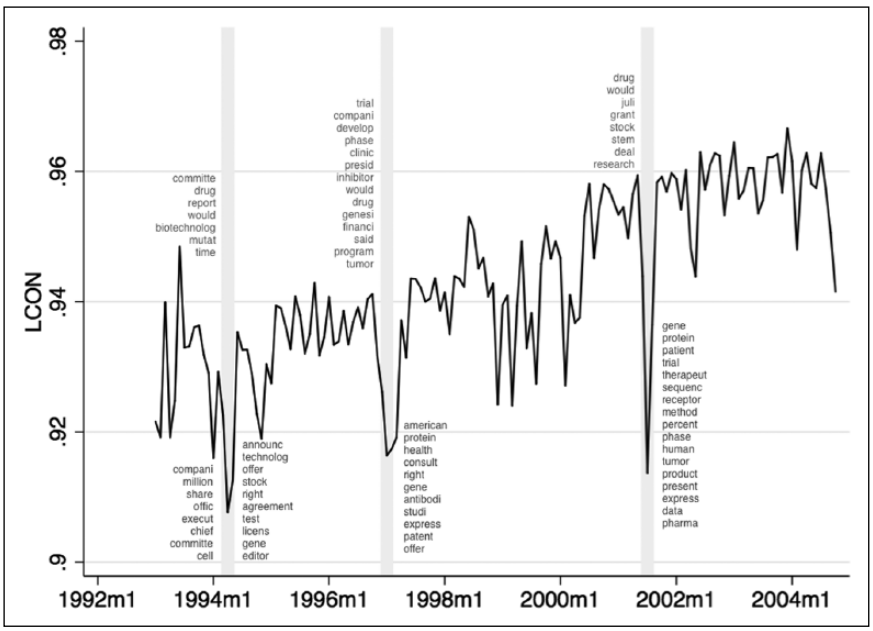

Figure 3. Highlighted deviations in monthly language consistency.

highlighted deviations occurred in March 1994, January 1997, and July 2001. Listed next to each deviation are the 25 word-stems, which contributed most significantly to the change in probability. The word-stems with positive change (increase in probability mass) are listed on top and those with negative changes (decreases in mass) are listed at the bottom.

A few patterns are apparent. The deviation in 1994 was led by more sharp declines than gains, with declines outpacing gains nearly 2.6 to 1. The declines suggest a waning attention to the early stages of biotechnology ventures, with the words *stock, share, chief, executive, announce*, and *agreement* all declining significantly. In contrast, the deviation in January of 1997 posts nearly even gains and declines, with the words *trial, phase, clinic*, and *drug* heralding a possible shift toward the logic of commercialization as promising therapies made their way through the difficult FDA approval process. This shift is further supported by declines in the words *study* and *patent*. The decline of the word *American* is also a signal of the globalization of the industry as previously mentioned.

The deviation in July of 2001 is again primarily driven by sharp declines in probability mass. While difficult to interpret from a reading of the word-stems, a review of the articles appearing around this time suggests a shift away from the science of biotechnology and toward the regulatory climate within which the industry was operating. This was shortly after George W. Bush started his first term as President of the United States and there was a concern among scientists that funding for stem cell research could be cut. This is supported by advances in the words *stem* and *grant*. There was also some discussion as to who would be approved to take over as head of the FDA. Although the word *approve* did not make to the top 25 in absolute value, it is in the top 20 words that made probability gains.

Discussion

As marketers, we are confronted by an increasingly complex world (Reibstein et al., 2009)—one characterized by markets that emerge and evolve at an unprecedented rate. This has led to an increase in the need for methods that can help researchers understand market dynamics. In the current work, I introduce one such method, and show how it can be applied to the emergence and evolution of the biotechnology industry during the 1990s and early 2000s. However, this method of analyzing the market can be applied in a much broader sense.

Consider, for instance, the practice of content marketing, email promotion, or digital advertising. In each of these efforts, the marketer is charged with developing text that captures the attention of a consumer (Moskowitz & Martin, 2008). While most practitioners still rely on a professional copy editor, many are beginning to supplement the creative process with A/B testing and/or machine learning (see, for example, Whitaker, 2017). When implementing an A/B test, the researcher is typically tasked with changing specific words or phrases to see which garners the most favorable response. In machine-learning applications, the researcher is often tasked with selecting the set of features upon which a model is trained. In both cases, the choice of features is arbitrary and based largely on the intuition and skill of the researcher.

To supplement this intuition, some have turned to software tools that provide recommendations. Indeed, Toubia and Netzer (2016) introduce one such tool based on a text's prototypicality. I suggest that recommendations based on language consistency can be equally productive. Marketing content needs to be at once novel and familiar. If there is no novelty, the consumer loses interest, but if a piece of content is too novel, the consumer might be confused. Using a method like the one described in this article, a researcher could specify the desired degree of surprise (i.e., novelty), and words or phrases that match could be surfaced from a context-specific corpus. Moreover, machine-learning algorithms can be trained on their adherence to some preferred level of novelty in the text.

Another potential application is to use language consistency as a means to spot emerging trends and market discontinuities. Social listening—the practice of using software to monitor online conversations—is now a major component of the modern marketer's research toolbox. Indeed, the

field is now crowded with vendors like Brandwatch, Sprout Social, Mention, Datorama, and LexisNexis (and many others), which help researchers evaluate brand sentiment, mentions, and share of voice. While these metrics have become a sort of industry standard, their value as a source of marketing intelligence is suspect.

For instance, it is often the case that promotion (whether traditional, social, or otherwise) is designed to change the minds of existing and potential customers. Sentiment is one important dimension of this change, but another is the actual change in discourse—that is, changes in the way customers are talking about the brand. By looking at changes in the consistency of language use (as compared to a pre-campaign baseline for instance), market researchers can quantify the amount of change that actually occurred as the result of their campaign. Moreover, this type of metric could be incorporated rather seamlessly into existing social listening platforms. When the language used by consumers changes enough, software could trigger an alert containing the words and phrases that created the discontinuity.

Finally, the concept of language consistency (and quantifying surprise more generally) can be used to aid the practice of qualitative research (Schmidt, 2010). Although some researchers argue that computer-assisted qualitative data analysis is of limited use in the generation of deep consumer insights (Dolan & Ayland, 2001), many others argue for its value (Humble, 2015). Regardless, there is growing evidence that computer–human partnerships are on the rise in qualitative market research as evidenced by the growing use of software like ATLAS, MAXQDA, and NVivo.

Nonetheless, the currently available tools are geared toward data management and keyword-based search—an approach that presumes the researcher is relatively aware of what they are looking for (Humble, 2015). From a methodological standpoint, this can lead to problems with researcher bias (as noted in the introduction of this article). However, there is also the practical issue of dealing with the massive quantities of (textual) data available to the modern market researcher—it is simply not reasonable to expect humans to read every piece of text available.

How, then, do qualitative researchers know where to start looking? It is often the stated desire of market-oriented ethnographers to identify evidence of both complementary and discrepant behavior in their data (Arnould & Wallendorf, 1994; Geertz, 1973). By using the method presented herein, one could surface language that is both consistent and inconsistent with the broader average. Such text would serve as a point of entry for the discovery of diverse behaviors. Note, for instance, that the major drops in language consistency in the biotechnology database corresponded with major events of interest. Analysis of such discrepant behavior can facilitate the oft sought-after "thick description" of marketing phenomena prized by qualitative researchers (Arnould & Wallendorf, 1994, p. 599).

Limitations to the current approach include a heavy reliance on category-specific textual data. For instance, data at the firm or even category level are often too sparse for analysis of the sort proposed. However, this is an area that would benefit from additional research—especially as larger and more robust datasets become available. Additional work could also look at more sophisticated features of the language. In the current work, features are constituted by an unstructured "bag" of individual word-stems. Future research could look at when word co-occurrence, longer phrases, or grammatical construction are important.

There is little doubt that analysis of big datasets of unstructured text shows great promise as a tool for social scientists and market researchers specifically (Klingenstein et al., 2014; Murdock et al., 2015). Yet, the science linking these data to our most important constructs is still in its infancy. These findings are one small step toward making this link a reality, and perhaps more importantly toward making these methods more applicable in practice.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Note for instance that compound words are split into their constituent parts so that a word like "don't" becomes three elements: "don," "'," and "t."
2. Note that many topic models can be viewed as a form of principal component analysis given a matrix of articles by terms (Blei, 2012).

References

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the internet. *Glottometrics*, *3*(1), 143–150.

Arnould, E. J., & Wallendorf, M. (1994). Market-oriented ethnography: Interpretation building and marketing strategy formulation. *Journal of Marketing Research*, *31*, 484–504.

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. New York, NY: Free Press.

Bateson, G. (1972). *Steps to an Ecology of Mind*. Chicago, IL: University of Chicago Press.

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

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). New York, NY: ACM.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*, 17–35.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text Mining: Classification, Clustering, and Applications*, *10*(71), 34.

Cattani, G., Ferriani, S., Negro, G., & Perretti, F. (2008). The structure of consensus: Network ties, legitimation, and exit rates of us feature film producer organizations. *Administrative Science Quarterly*, *53*, 145–182.

Dolan, A., & Ayland, C. (2001). Analysis on trial. *International Journal of Market Research*, *43*, 377–389.

Dolbec, P. Y., & Fischer, E. (2015). Refashioning a field? Connected consumers and institutional dynamics in markets. *Journal of Consumer Research*, *41*, 1447–1468.

Geertz, C. (1973). *The interpretation of cultures* (Vol. 5019). Basic Books.

Goldberg, A. (2012). *Where do social categories come from? A comparative analysis of online interaction and categorical emergence in music and finance* (PhD thesis). Princeton University, Princeton, NJ.

Green, S. E., Li, Y., & Nohria, N. (2009). Suspended in self-spun webs of significance: A rhetorical model of institutionalization and institutionally embedded agency. *Academy of Management Journal*, *52*, 11–36.

Hakala, H., Nummelin, L., & Kohtamäki, M. (2017). Online brand community practices and the construction of brand legitimacy. *Marketing Theory*, *17*, 537–558.

Hardy, C. (2011). How institutions communicate; or how does communicating institutionalize? *Management Communication Quarterly*, *25*, 191–199.

Harmon, D. J., Green, S. E., & Goodnight, G. T. (2015). A model of rhetorical legitimation: The structure of communication and cognition underlying institutional maintenance and change. *Academy of Management Review*, *40*, 76–95. doi:10.5465/amr.2013.0310

Humble, Á. (2015). Review essay: Guidance in the world of computer-assisted qualitative data analysis software (CAQDAS) programs. In: *Forum: Qualitative social research* (Vol. 16). Berlin, Germany: Freie Universität Berlin.

Humphreys, A. (2010). Megamarketing: The creation of markets as a social process. *Journal of Marketing*, *74*(2), 1–19.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall PTR.

Klingenstein, S., Hitchcock, T., & DeDeo, S. (2014). The civilizing process in London's old bailey. *Proceedings of the National Academy of Sciences*, *111*, 9419–9424.

Lawes, R. (2002). Demystifying semiotics: Some key questions answered. *International Journal of Market Research*, *44*, 251–265.

Lounsbury, M., & Crumley, E. T. (2007). New practice creation: An institutional perspective on innovation. *Organization Studies*, *28*, 993–1012.

Mele, C., Pels, J., & Storbacka, K. (2015). A holistic market conceptualization. *Journal of the Academy of Marketing Science*, *43*, 100–114.

Moskowitz, H. R., & Martin, B. (2008). Optimising the language of email survey invitations. *International Journal of Market Research*, *50*, 491–510.

Murdock, J., Allen, C., & DeDeo, S. (2015). Exploration and exploitation of Victorian science in Darwin's reading notebooks. arXiv preprint arXiv:150907175.

Peñaloza, L., & Venkatesh, A. (2006). Further evolving the new dominant logic of marketing: From services to the social construction of markets. *Marketing Theory*, *6*, 299–316.

Petkova, A. P., Wadhwa, A., Yao, X., & Jain, S. (2014). Reputation and decision making under ambiguity: A study of us venture capital firms' investments in the emerging clean energy sector. *Academy of Management Journal*, *57*, 422–448.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Powell, W. W., & Colyvas, J. A. (2008). Microfoundations of institutional theory. In R. Greenwood, C. Oliver, K. Sahlin, & R. Suddaby (Eds.), *The SAGE handbook of organizational institutionalism* (pp. 276–298). London, England: SAGE.

Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences1. *American Journal of Sociology*, *110*, 1132–1205.

Reibstein, D. J., Day, G., & Wind, J. (2009). Guest editorial: Is marketing academia losing its way? *Journal of Marketing*, *73*(4), 1–3.

Rolland, S. E., & Parmentier, G. (2013). The benefit of social media. *International Journal of Market Research*, *55*, 809–827.

Rosa, J. A., Porac, J. F., Runser-Spanjol, J., & Saxon, M. S. (1999). Sociocognitive dynamics in a product market. *Journal of Marketing*, *63*, 64–77.

Santos, F. M., & Eisenhardt, K. M. (2009). Constructing markets and shaping boundaries: Entrepreneurial power in nascent fields. *Academy of Management Journal*, *52*, 643–671.

Schmidt, M. (2010). Quantification of transcripts from depth interviews, open-ended responses and focus groups: Challenges, accomplishments, new applications and perspectives for market research. *International Journal of Market Research*, *52*, 483–509.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, *20*, 571–610.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, *51*, 463–479.

Toubia, O., & Netzer, O. (2016). Idea generation, creativity, and prototypicality. *Marketing Science*, *36*(1), 1–20.

Tripsas, M. (2009). Technology, identity, and inertia through the lens of "the digital photography company." *Organization Science*, *20*, 441–460.

Vinhas, A. S., Chatterjee, S., Dutta, S., Fein, A., Lajos, J., Neslin, S., . . .Wang, Q. (2010). Channel design, coordination, and performance: Future research directions. *Marketing Letters*, *21*, 223–237.

Whitaker, A. (2017). *Artificial intelligence makes for savvy content marketers*. Retrieved from https://www.forbes.com/sites/forbesagencycouncil/2017/10/06/artificial-intelligence-makes-for-savvy-content-marketers/

Wolff, G. (2001). *The biotech investor's bible*. New York, NY: John Wiley & Sons.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Zucker, L. G. (1983). Organizations as institutions. *Research in the Sociology of Organizations*, *2*(1), 1–47.