Some Comments on
"The Effect of Calorie Posting Regulation on Consumer Opinion"
by Dinesh Puranam, Vishal Narayan, and Vrinda Kadiyali

Gerry Katz
Applied Marketing Science, Inc.
Waltham, MA

December 15, 2017

Overview

This paper brings together two important research interests in the field of marketing science, one having been around since its earliest days and one having emerged only in the past few years. The former is the ability to measure the effect of new information on consumer behavior, while the latter is the ability to apply machine learning to social media content. The two come together beautifully in this paper.

Discussion

As students in the natural sciences, we all learn about the *scientific method* at an early age. We observe phenomena, form hypotheses, and conduct experiments in order to confirm or deny those hypotheses. Experiments generally use the concept of a *test* cell to be compared to a *control* cell. That is, we conduct a test in which we alter one variable (the dependent variable) and measure the impact on another variable (the independent variable), all while attempting to isolate and hold constant everything else in the experiment. It usually works quite well in the natural sciences.

But when it comes to marketing science, we run into all kinds of difficulties. We think to ourselves, shouldn't it be easy to just increase or decrease advertising and measure its impact on sales? Very early in my career, I learned just how difficult that is. There was so much noise from confounding effects such as promotion, price changes, regional differences, changes in distribution, and competitive actions. Then there were all of the measurement issues. How do we know who saw the ads, and who made the purchases? We quickly learned just how complicated it is to link marketing actions to sales response.[1] Furthermore, because of the difficulty and expense involved in conducting formal experimentation, we often had to rely on what my mentor, John Little,

called *naturally occurring experiments* – periods in which we knew of a change in some independent variable and tried to observe its impact on sales. But because of the difficulty of linking and measuring these effects, we often had to focus on an intermediate variable – such as impact on customer attitude or opinion – rather than choice as the dependent variable.

Such is the case with *"The Effect of Calorie Posting Regulation on Consumer Opinion"* by Puranam, Narayan, and Kadiyali. In this paper, they attempt to examine the impact of Mayor Michael Bloomberg's 2008 regulatory change in New York City requiring that all chain restaurants with 15 or more outlets post calorie information on their menus. Would this change make calories salient inside the restaurant, potentially redirecting consumers to healthier foods and/or restaurants?

To examine this problem, they relied heavily on a second research interest in the field of marketing science – the explosion of massive amounts of virtually free data on the Internet – what we now refer to as *Big Data*. Most of this research attempts to apply quantitative analysis on what is mostly transaction-level data (clicks, searches, purchases, etc.). But more recently, computer scientists have turned their attention to the application of artificial intelligence (AI) on textual information – what we now call *Text Analytics* on *User-Generated Content* (UGC), i.e., the ability of computers to glean knowledge from the examination of natural language (text) beyond what the machine has explicitly been told to look for.

That is the essence of this paper. The user-generated content, in this case, is restaurant reviews posted on all kinds of websites. Since these reviews have become so prevalent, Puranam, Narayan, and Kadiyali focus their analysis on the content of these reviews. Using a probabilistic construct from the field of computer science called Latent Dirichlet Allocation (LDA) modeling, they examine the effect of this extremely interesting "naturally occurring experiment" in which, at a specific point in time (early 2008), certain kinds of restaurants (chains with 15 or more outlets) were required to post calorie counts on each of their menu items, while others (standalone restaurants) were not.

So the test-versus-control scenario is defined as *chain* restaurants (the test cell) versus *standalone* restaurants (the control cell) as observed in the period *preceding* the regulatory change starting in 2008 versus the period *following* the regulatory change. In both cases, the dependent variable is the proportion of review content dealing with health.

In their analysis, the “machine” reads all reviews of restaurants from 2004 to 2012, deciding what words are most closely related to one another (an affinity diagram of sorts), and hence, are assumed to define a “topic”. For example, the machine clustered the words: *calories, fat, healthy, menu, count, low, muscle,* etc. into one topic that they labeled *Health.* Other topics identified by this process included *Brand, Service, Food, and Hygiene*. For completeness, they then further “seeded” the *Health* topic with a number of other key words that were likely to be associated with that topic. Except for this seeding process, the algorithm made all the decisions as to what is and is not meaningful discussion belonging to each topic.

The analysis included content from nearly 762,000 reviews of New York City restaurants posted between late 2004 and the end of 2012. It covered about 9,800 different restaurants and 77 unique chains. The length of the average review was 126.7 words, and across all of the reviews, there were 44,276 unique words (not counting so-called stop words like “a” and “the”). In the end, they were able to conclude that there was a statistically small but significant increase in the proportion of restaurant review content dealing with health issues. And this increase could be attributed largely to reviewers who had never posted before. They go on to address nearly a dozen potential challenges to the robustness of their conclusion, none of which were able to overturn it.

Summary

This paper touches on two important aspects of both marketing and computer science: the impact of new information on consumer behavior, and the ability of computers to deal intelligently with human language, not just numbers. This latter issue is truly a new frontier in all of the social sciences, and I suspect that we are only at the beginning.

---

[1] One of the most successful attempts to deal with these problems in the Consumer Package Goods industry was the BehaviorScan® system created by information Resources, Inc. in the 1980s. In this highly creative “laboratory”, they selected eight small cities in which they could arrange to control television advertising to various populations through the cable TV system and link it to actual purchases by those populations using supermarket scanners and personal identification cards.