Title: Distinguishing reviews about conventional and alternative medicine using textual analysis

Abstract

Despite the success of modern, conventional medicine in treating disease, alternative medical establishments continue to flourish. Do patients/patrons of these two kinds of medicine view their treatment experiences differently? If so, what factors influence these differing views? Customer reviews of medical establishments in Canada, Germany, the United Kingdom, and the United States showed that patients/patrons rate alternative medical establishments more highly than conventional medical establishments. In the reviews, the numbers of words, characters, syllables, punctuation marks, and pronouns; polarity; diversity; lexical classification; and formality did not distinguish between conventional and alternative medicine. But the usage rates of the words "massage" and "chiropractor" did accurately classify reviews with ~90% accuracy. Further analysis of the usage rates of other words may yield insights into why reviewers rate alternative medicine more highly than conventional medicine.

Introduction

In recent centuries, conventional medicine has developed practices to treat disease that are largely based on scientific evidence, and it has developed professional training and licensing standards to ensure that medical treatments are delivered safely and effectively. Practitioners of alternative medicine – such as naturopathy and reiki – also claim to treat disease and improve health, even though their practices are not currently supported by robust scientific evidence.

Patients, or patrons, of conventional and alternative medicines have similar goals of alleviating disease and improving health. But do these patients/patrons have different perceptions of their treatment experiences with each kind of medicine? And if so, what drives these different perceptions? Any such differences might arise from different types of patients/patrons, different goals for treatment, different treatment results, different "customer" service (e.g., "bedside manner"), or other factors.

Analyzing customer reviews of medical businesses/establishments might provide insight into these questions. The company Yelp hosts a website where customers may rate and review many types of businesses, including conventional and alternative medical establishments. Yelp reviews were analyzed to provide insight into the questions posed above. More specifically, do patients/patrons of each kind of medicine rate their treatment experiences differently? Can textual analysis of the reviews of medical establishments predict whether the review was written about conventional or alternative medicine? If so, the predictive factors may illuminate how the treatment experiences differ between conventional and alternative medicine. Such information might be useful to those who run medical establishments; such persons may wish to understand why patrons find their own establishment, or competing establishments, appealing. Public policy makers, healthcare investors, and researchers of the healthcare system might also find the results of this analysis interesting.

Methods

Data source. Yelp reviews were obtained from a publicly available data set that included reviews of businesses in 10 cities in Canada, Germany, the United Kingdom, and the United States (http://www.yelp.com/dataset_challenge (http://www.yelp.com/dataset_challenge)).

Identifying reviews of conventional and alternative medicine. Conventional and alternative medical establishments were identified by searching for a broad set of business category terms that were associated with each type of medicine and then eliminating any businesses that were categorized with any terms associated with the other type of medicine (see Appendix 1 for details). This approach ensured that any establishments that were associated with both conventional and alternative medicines were not included in the analysis, so that the remaining establishments were clearly and unambiguously associated with only one type of medicine.

Once these conventional and alternative medical establishments were identified, reviews of these establishments were identified. Any reviews encoded in UTF-8 were eliminated because they were not readable by textual analysis functions. Two remaining reviews were not written in English and therefore were eliminated. One additional review was eliminated because it produced an error for undetermined reasons.

Ratings analysis. Each Yelp review had an associated rating of 1 - 5 stars (1 = lowest rating of the business; 5 = highest). These ratings were analyzed using a chi-square test of independence (R function ‘chisq.test’).

Features / Predictor variables. Functions from the R package ‘qdap’ were used to analyze the text of each review. The final feature set for each review included the numbers of words, characters, syllables, and poly-syllables; the numbers of characters, syllables, and poly-syllables per word; the numbers of periods, question marks, exclamation marks, and incomplete statements; the numbers of pronouns (i.e., “I”, “you”, “he”, “she”, “they”, “it”, and “me”); polarity; diversity (Shannon, Simpson, and Berger-Parker indices); lexical classification; and formality. The number of sentences (and variables based on this metric) was excluded as a feature because the function measuring sentence number relies on punctuation consistent with formal English, which was not reliably present in Yelp reviews. The Renyi index of diversity was eliminated as a feature because it correlated so highly with other measurements of diversity.

Additionally, the usage rates of words appearing in reviews were included as features. Because the sentence number was not reliably measurable, the usage rate for a word in a review was expressed per total number of words in the review. Words that referred to particular people or locations were eliminated so that analysis results would better generalize to other data sets with reviews from other medical establishments and/or locations.

To limit computation time, the final feature set included only usage rates of the words most likely to discriminate between reviews for conventional and alternative medicine. To determine which words would be the best discriminators, the differences in usage rates between conventional and alternative medicine reviews (expressed as a proportion of the usage rate in conventional medicine reviews for that word) were calculated. These differences were multiplied by the word’s frequency (where frequency in alternative medicine reviews was over-weighted to correspond to machine learning model training; see ‘Machine learning prediction’ below), and words with the largest values were deemed likely to be the best discriminators. Thus, words that were used most frequently and that had the largest usage rate differences between conventional and alternative medicine reviews were considered likely to be the best discriminators.

In addition to features derived from the text of the reviews, Yelp users’ ratings of each review were included as features. These ratings were votes on whether the review was funny, useful, and “cool”.

Machine learning prediction. The reviews were divided so that 60% were assigned to a training data set and the remainder were assigned to a testing set. These assignments were stratified by star ratings, so that approximately equal proportions of each rating (1 - 5 stars) were in the training and testing sets.

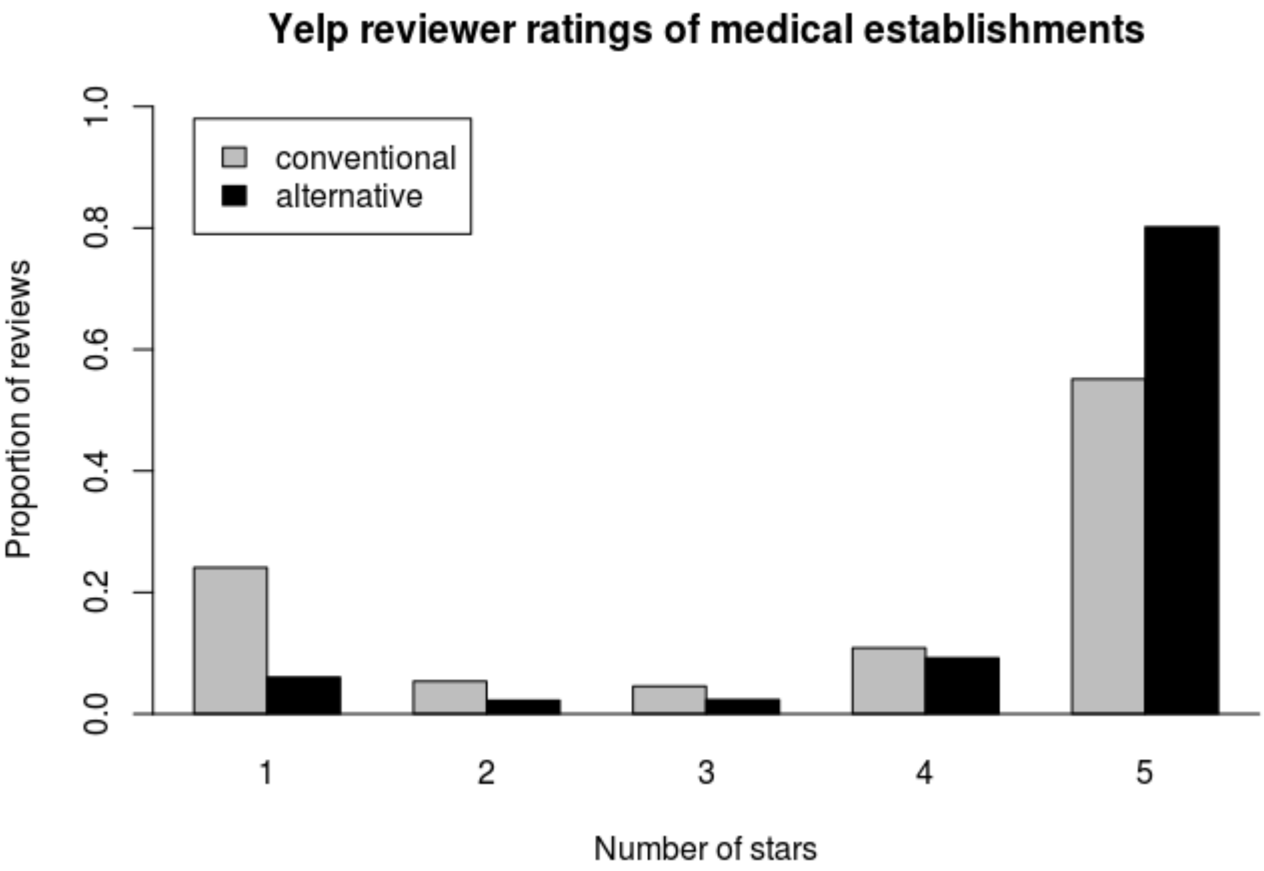
Because there were many fewer reviews of alternative medicine establishments compared to conventional medicine establishments, reviews for alternative medicine were oversampled with stratification by star rating, so that there were equal numbers of reviews for conventional and alternative medicine.

A classification tree function (‘rpart’ in the R ‘caret’ package) was selected because classification trees provide more interpretable results than many other machine learning algorithms. The tree was trained on the training set and then tested once on the testing set. A confusion matrix was computed for examination.

Results

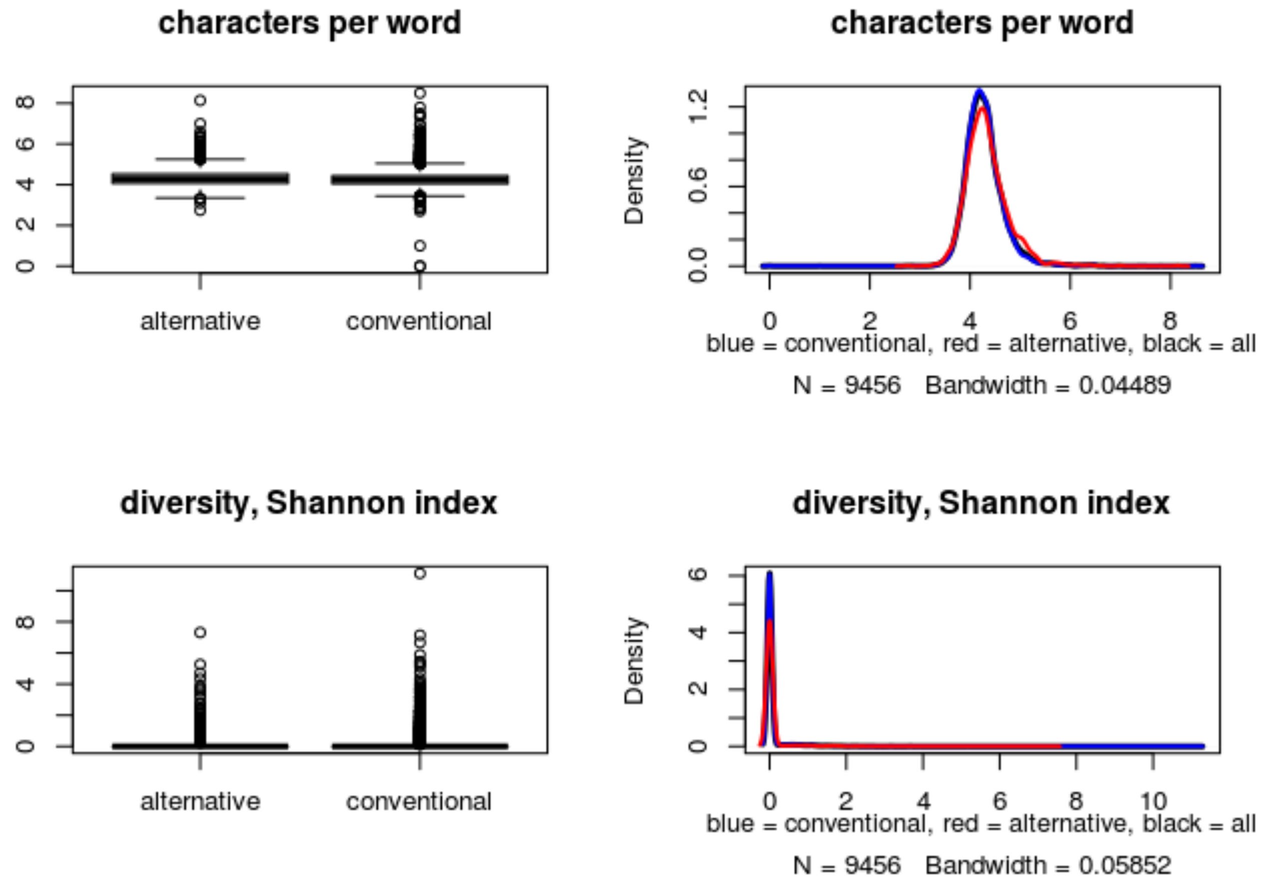
For analysis, 12895 reviews of 1847 conventional medical establishments and 2866 reviews of 368 alternative medical establishments were identified.

I first investigated whether Yelp reviewers rated their treatment experiences differently between conventional and alternative medicine. The proportions of reviews receiving 1, 2, 3, 4, or 5 stars for each kind of medicine are below.



A chi-square test of independence suggested that the star ratings are not independent of the type of medicine, chi-squared = 695.9389492, df = 4, p = 2.639761910^{-149}, alpha = 0.05. In other words, the distribution of ratings differs according to the type of medicine. Alternative medicine establishments appear to receive a higher proportion of the highest rating (5 stars) than conventional medicine establishments; correspondingly, conventional medicine receives a higher proportion of the lowest rating (1 star). In other words, reviewers tend to rate alternative medicine establishments higher than conventional medicine establishments.

Next, I examined exploratory graphs to better understand which features might distinguish between reviews of conventional and alternative medicine. Example graphs for “characters per word” and diversity, as measured by the Shannon index, are below.



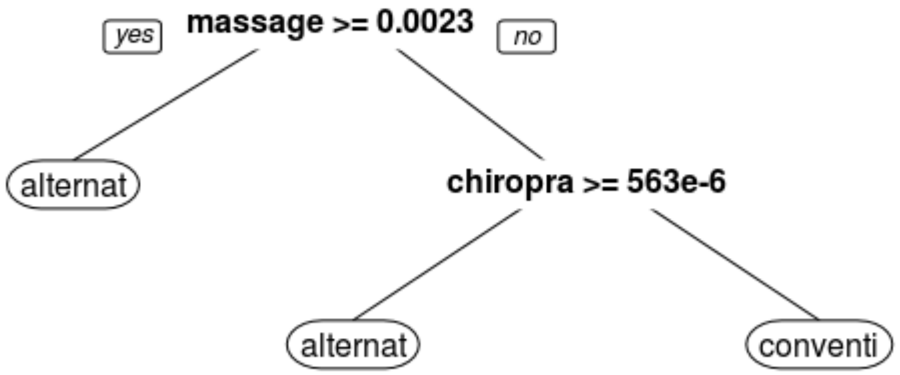
The exploratory graphs did not identify any features that clearly distinguished between reviews of conventional and alternative medicine. The plots of the two features above (“characters per word” and “diversity, Shannon index”) showed some of the largest differences of any of the potential predictor variables. Even so, the exploratory graphs are univariate, and a machine learning algorithm might identify more complex relationships among combinations of features to enable a better distinction between conventional and alternative medicine.

Below are the words deemed most likely to be the best discriminators between reviews for conventional and alternative medicine. They are listed in order from most to least discriminatory.

```
##  [1] "acupuncture"   "masseuse"      "massage"       "chiropractor" 
##  [5] "massages"      "chiropractors" "adjustment"    "adjustments"  
##  [9] "neck"          "body"          "therapist"     "adjusted"     
## [13] "tissue"        "therapists"    "spa"           "joint"        
## [17] "feet"          "relaxing"      "shoulders"     "foot"         
## [21] "session"       "healing"       "therapy"       "lower"        
## [25] "tip"           "pain"          "back"          "wellness"     
## [29] "shoulder"      "energy"        "strong"        "relaxed"      
## [33] "headaches"     "price"         "pressure"      "relief"       
## [37] "treatments"    "spine"         "relax"         "feeling"      
## [41] "hip"           "hot"           "quiet"         "areas"        
## [45] "pains"         "chronic"       "dentist"       "community"    
## [49] "table"         "deep"          "beat"          "muscle"       
## [53] "helped"        "clients"       "better"        "place"        
## [57] "office"        "injury"        "approach"      "sore"         
## [61] "dental"        "accident"      "atmosphere"    "sessions"     
## [65] "affordable"    "car"           "rough"         "full"         
## [69] "owner"         "music"         "they"          "teeth"        
## [73] "physical"      "prices"        "help"          "places"       
## [77] "doctor"        "effective"     "gift"
```

A classification tree algorithm on the training set (60% of the data) yielded the model plotted below.

```
## n= 15474 
## 
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
## 
## 1) root 15474 7737 alternative (0.500000000 0.500000000)  
##   2) massage>=0.00233808 2812   31 alternative (0.988975818 0.011024182) *
##   3) massage< 0.00233808 12662 4956 conventional (0.391407361 0.608592639)  
##     6) chiropractor>=0.0005630631 839    5 alternative (0.994040524 0.005959476) *
##     7) chiropractor< 0.0005630631 11823 4122 conventional (0.348642477 0.651357523) *
```



Reviews that more frequently mentioned “massage” or “chiropractor” were classified as pertaining to alternative medicine.

The predictions of this classification tree on the testing set (the remaining 40% of the data) are below.

```
##               Reference
## Prediction     alternative conventional
##   alternative          501           27
##   conventional         646         5131
```

This classification tree yielded an accuracy of 0.8932593. As a comparison baseline, a random classifier would be expected to have an accuracy of about 50%.

Discussion

The 0.8932593 accuracy rate by the classification tree was a substantial improvement over chance (50%) in distinguishing between reviews pertaining conventional or alternative medicine. The tree incorporated only two predictors to achieve this accuracy: the appearance of “massage” or “chiropractor” at high-enough rates predicted that the review was about alternative medicine instead of conventional medicine.

The classification tree did achieve substantial accuracy, but did not provide much insight into how patients/patrons of medical establishments might perceive their experiences differently between conventional and alternative medicine. Instead, the tree’s predictors simply mention a service (“massage”) and practitioner (“chiropractor”) associated with alternative medicine without providing further insight.

Yelp reviewers rated alternative medicine establishments higher than conventional medicine establishments. Yelp reviewers might have better experiences with alternative medicine for any number of reasons. For example, they might harbor pre-existing biases favoring alternative medicine. Or they might choose to treat less severe problems with alternative medicine and more severe, less tractable problems with conventional medicine. Or alternative medical practitioners might provide better resolution of a medical problem or maybe better customer service in ways not directly related to a medical problem. Regarding customer service, it is notable that among the words that might best discriminate between reviews of alternative and conventional medicine were "price", "affordable", "community", "atmosphere", and "music". Or, it is possible that patients/patrons are predisposed to rate alternative medicine higher because their purposes for the two kinds of medicine differ; perhaps they more often use alternative medicine to pursue positive health states instead of avoiding negative health states, for which they more often use conventional medicine. In this regard, the words "relax" and "relaxed" appeared about 7 times more often in reviews for alternative medicine than for conventional medicine. While many other explanations might explain the difference in ratings between conventional and alternative medical establishments, the cursory speculation here suggests that a more thorough analysis of the word frequencies used in reviews might provide further insights.

An earlier version of this report was not able to incorporate analysis using word usage rates. Instead, only the numbers of words, characters, syllables, punctuation marks, and pronouns; polarity; diversity; lexical classification; and formality (as described above in "Features / Predictor variables") were included. Notably, the classification tree for that analysis did no better than chance (50%) in classifying reviews, suggesting that none of those variables differed usefully between reviews for conventional and alternative medicine. More sophisticated machine learning techniques, like random forests, might discover more complex relationships among those variables that were overlooked by the classification trees implemented here.

While word usage rates were always intended to be included in this analysis, deadline pressures prevented their inclusion in the earlier version of this report. Additionally, the sample size of alternative medicine reviews was too small to support division of the data set into training, testing, and validation subsets, so it was divided only into training and testing data. Consequently, the testing data was used for validation twice – the second time with a model that was developed after the first validation. This could theoretically lead to overfitting of the classification tree in this report, so that the tree's classification accuracy of about 90% could be inflated. Because the word usage rates were always intended for inclusion and because the variable selection was not otherwise finely tuned, the risk of overfitting is probably minimal, but it is present.

Notably, this analysis included only reviews of establishments that could clearly be classified as conventional medicine or alternative medicine. Establishments, such as yoga studios or gyms, that might be compatible with both kinds of medicine were excluded. This approach was chosen to accentuate the differences between the two kinds of medicine and make prediction success more likely, but it might also obscure the true relationship between them.

Appendix 1

Conventional medicine establishments were identified by first searching for all businesses categorized by the following terms: Doctor, Hospital, Allergist, Anesthesiologist, Cardiologist, Surgeon, Dentist, Drugstore, Ear Nose & Throat, Endodontist, Internal Medicine, Laser Eye Surgery/Lasik, Obstetrician, Gastroenterologist, Gynecologist, Ophthalmologist, Oncologist, Orthodontist, Orthopedist, Orthotic, Pediatric, Periodontist, Pharmacy, Podiatrist, Psychiatrist, Pulmonologist, Radiologist, Rheumatologist, Urologist, and Medical Center. These terms were chosen to include a broad selection of conventional medical establishments. From the resulting list of businesses, all businesses that were categorized by terms associated with alternative medicine (or terms not clearly associated with conventional medicine) were eliminated. These terms included: Acupuncture, Massage, Naturopath, Psychic, Yoga, Spas, Food, Fitness, Osteopath, Chinese Medicine, Shopping, Chiropractor, Cannabis, Reflexology, Rolfing, Coach, and Reiki.

Alternative medicine establishments were identified by first searching for all businesses categorized by the following terms: Acupuncture, Chiropractor, Chinese Medicine, Reflexology, Reiki, Osteopath, Rolfing, and Naturopathic. These terms were chosen to include a broad selection of alternative medical establishments. From the resulting list of businesses, all businesses that were categorized by terms associated with conventional medicine (or terms not clearly associated with alternative medicine) were eliminated. These terms included: Dermatologists, Neurologist, Obstetrician, Gynecologist, Orthopedist, Allergist, and Internal Medicine.