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

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. Four additional reviews were eliminated because they produced errors 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. After features with too little variability were removed (by the 'nearZeroVar' function in the 'caret' R package), 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.

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.

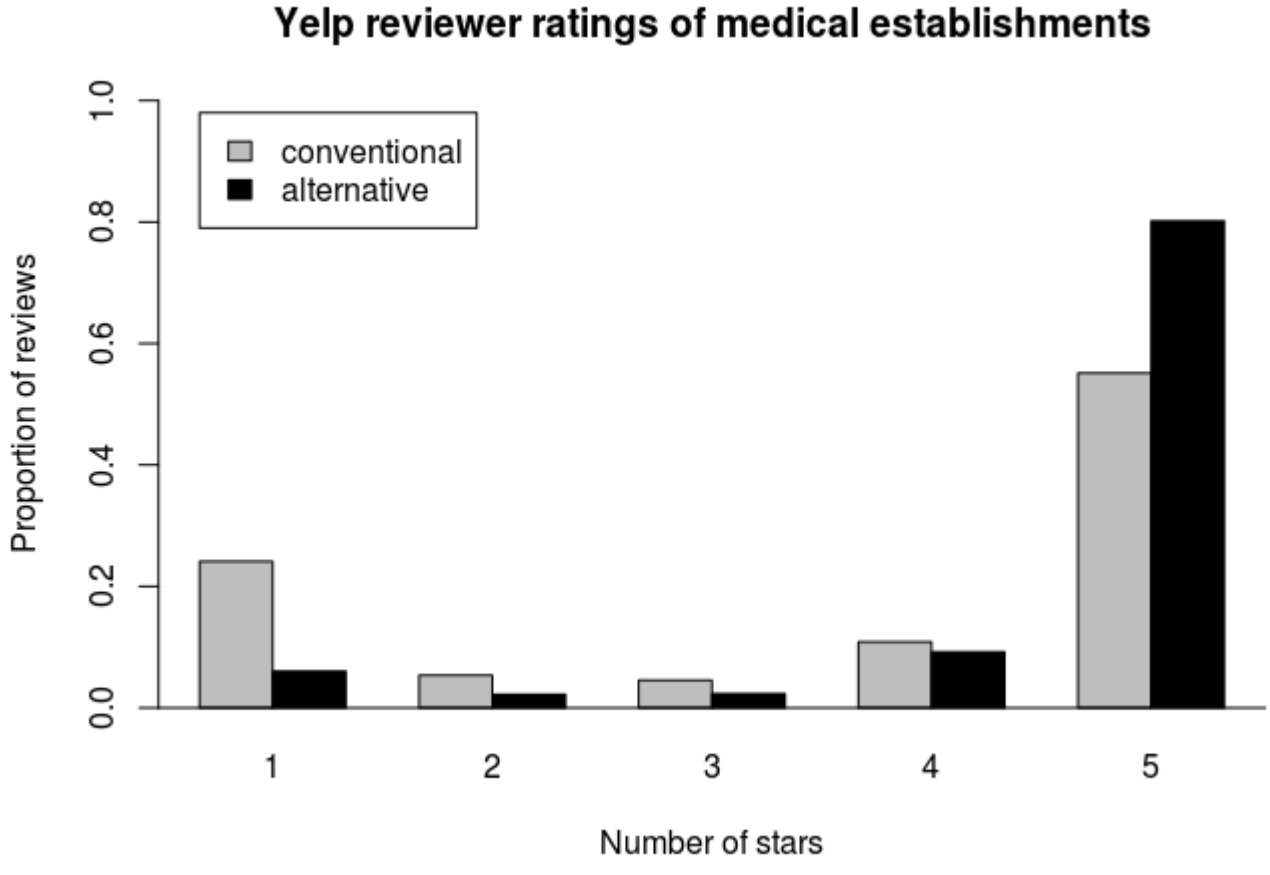
Because there were many fewer reviews of alternative medicine establishments compared to conventional medicine establishments, reviews for alternative medicine were oversampled with stratification by star rating, so that there were equal numbers of reviews for conventional and alternative medicine.

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

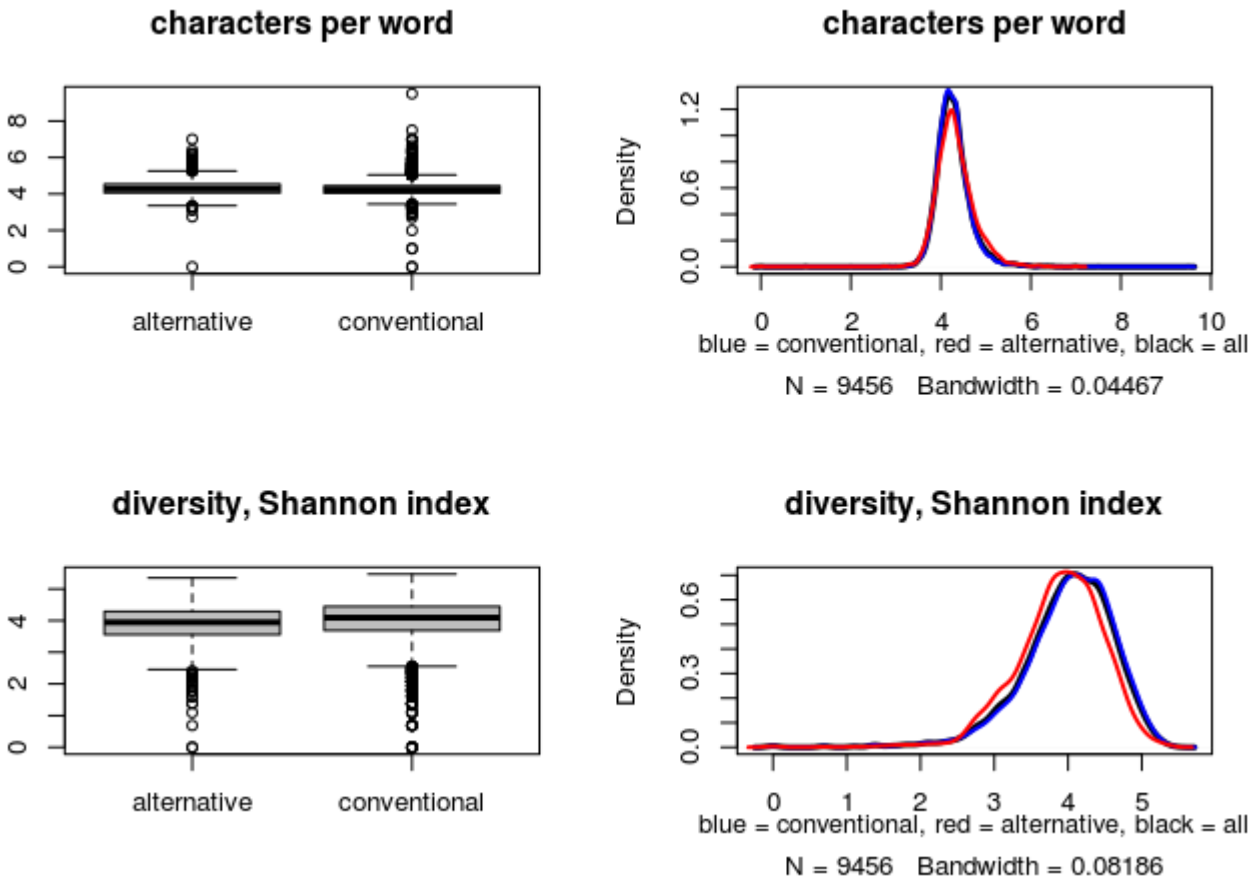
For analysis, 12894 reviews of 1848 conventional medical establishments and 2866 reviews of 368 alternative medical establishments were identified.

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.



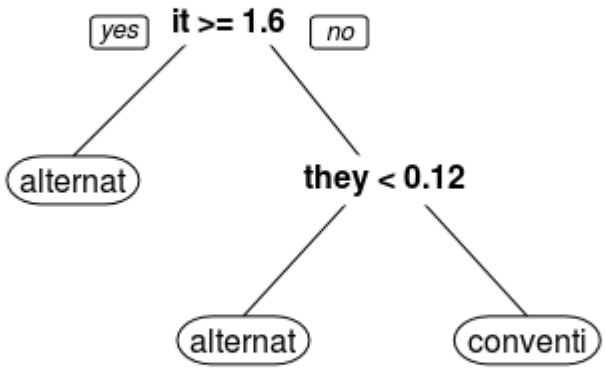
A chi-square test of independence suggested that the star ratings are not independent of the type of medicine, chi-squared = 696.466185, df = 4, p = 2.029574110^{-149}, alpha = 0.05. In other words, the distribution of ratings differs according to the type of medicine. Alternative medicine establishments 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.

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



Reviews that contained the pronoun “it” at least 1.6 times were classified as pertaining to alternative medicine. Reviews with fewer than 1.6 “it”s were classified for alternative medicine if the pronoun “they” was used less than 0.12 times; otherwise, they were classified for conventional medicine.

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

```
##               Reference
## Prediction     alternative conventional
##   alternative          844         2831
##   conventional         303         2326
```

This classification tree yielded an accuracy of only 0.5028553. As a comparison baseline, a random classifier would be expected to have an accuracy of about 50%.

Discussion

The classification tree did not succeed in distinguishing between reviews pertaining conventional or alternative medicine. The tree’s prediction accuracy of only 0.5028553 was not substantially better than random classification, which would provide about 50% accuracy.

One possible explanation for the classification tree model’s poor performance is that reviewers of conventional and alternative medicine establishments use similar language in describing their treatment experiences. Patients/patrons might focus on similar aspects of their treatment experience, like bedside manner and waiting times. Even if conventional medicine is more successful in treating ailments than alternative medicine, conventional treatment limitations, side effects, and placebo effects might obscure detecting any such successes in patient reviews. Patients/patrons did review alternative medicine somewhat more highly than conventional medicine, but the machine learning procedure controlled for ratings and therefore could not use them to distinguish between conventional and

alternative medicine reviews.

Another possible explanation is that the most useful features for distinguishing reviews of alternative or conventional medicine were not extracted from the reviews. This report's deadline limited further efforts to improve prediction accuracy, but the next step would be to include frequencies of the most commonly used words in the reviews. For example, alternative medicine reviewers might write more about "relaxation" whereas conventional medicine reviewers might write more about "insurance companies." Such information might substantially improve prediction.

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. Many patients/patrons may not see the two kinds of medicine as all that different, and the results of this study can not reject that position.

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.