

COMPETITIVE ANALYSIS OF ONLINE REVIEWS USING EXPLORATORY TEXT MINING

William J. Amadio
J. Drew Procaccino

Preliminary communication

Received 14 March 2016

Revised 21 June 2016

29 June 2016

Accepted 13 July 2016

<https://doi.org/10.20867/thm.22.2.3>

Abstract

Purpose – This paper explores the usefulness of analyzing text-based online reviews using text mining tools and visual analytics for SWOT Analysis, as applied to the hotel industry. These results can be used to develop competitive actions.

Design – The text mining/visualization tool, *ReviewMap*, was used to transform an archive of reviews spanning multiple suppliers into a hierarchy of data of increasing dimensionality. Visual summaries at each level were integrated to propagate selections at one level throughout the rest of the hierarchy. These visual summaries identify features required for competition at a given level and features that currently discriminate amongst competitors.

Methodology – The approach was exploratory, the objective of which was to determine if useable competitive intelligence could be found in a typical collection of online reviews for a set of competing hotels. A publically available collection of reviews was subjected to a set of text mining procedures and visual analyses in order to summarize the features and opinions expressed.

Originality – Prior analyses of online reviews relied solely upon numeric “star” ratings. This study utilized text mining to uncover information within the written comments and applied the information in a SWOT Analysis of three competing hotels.

Findings – In the set of reviews used in this paper, a common measure of analytical power almost doubled when text mining summaries of the written comments were used in combination with numeric ratings. Visual analytics revealed the dominant features for each hotel, the features required of all hotels competing at a given level, and the features that define specific positions within the competitive landscape. This analysis of strengths, weaknesses, opportunities and threats revealed several promising competitive actions for the hotels in the study.

Keywords text mining, online reviews, competitive analysis, visual analytics, ReviewMap, SWOT

INTRODUCTION

Online review sites such as *TripAdvisor* provide consumers with unprecedented power to find products and services that meet specific needs, as well as provide feedback to suppliers. This paper focuses on the online review site as a feedback channel. It is common to see suppliers respond to individual reviews in an effort to reinforce positive experiences or make amends for negative ones. This study expands the suppliers’ sights beyond their own reviews to those of the competition, thereby utilizing the online review archive as a source of competitive intelligence on strengths, weaknesses, opportunities, and threats (SWOT Analysis) (Köseoglu et al, 2016).

Accessing reviews for competitive intelligence through a review site's interface is impractical since such interfaces are designed for manual processing of one supplier at a time. A competitive analysis must come from an automated system that relies upon text mining tools to summarize an archive of reviews spanning multiple suppliers, and then to identify relationships. Modern visualization technologies are also required to present results in meaningful ways.

The term *social media analytics* is often used to describe the tasks and technologies included above, and several carefully-crafted models of social media analytics have been published recently (Zeng et al, 2010; Fan & Gordon, 2014). This paper utilizes *ReviewMap*, an integrated text mining and visualization system that is consistent with the Capture-Understand-Present Methodology (Fan & Gordon 2014). The authors' experiences with *ReviewMap* that are presented in this paper provide a basis for evaluating the capabilities of text mining tools to support realistic analyses.

The paper begins with a brief review of the social media analytics literature, concentrating on analytics for competitive intelligence. This review provides a specific focus within social media analytics. Next, a set of text mining and visualization techniques are described that have been shown to be effective for the social media analytics focus of this study. The chosen tools were then integrated into a package we call *ReviewMap* and the result applied to a sample from Wang (2010) for three New York City hotels. Lastly, the paper concludes with a discussion of results, design choices, and ideas for future work.

1. LITERATURE REVIEW

McFarland and Ployhart (2015) suggest that social media can be defined as “digital Web 2.0 platforms that facilitate information sharing, user-created content, and collaboration across people”. These platforms include any Web or mobile-based technology that facilitates “interactive dialogue between organizations, communities, and individuals”. Included content can be in the form of text, images, or recorded or streamed audio and video.

The focus of this paper is text-based analytics, which includes natural language processing and text mining (Ittoo, Nguyen and van den Bosch, 2016). Fan & Gordon (2014) presented social media analytics as a three-stage process: (1) *capture*, (2) *understand*, and (3) *present*. This paper focuses on the *understand* phase, which involves modeling the captured data and gaining insights from these models, and the *present* stage, which deals with displaying findings from the *understand* stage using visual analytics.

The competitive analyses utilized in this paper are SWOT Analysis (Köseoglu et al, 2016) which deals with strengths, weaknesses, opportunities and threats, and customer segmentation. This type of analysis is usually discussed in terms of variables such as benefits sought, product feature preference, product usage, and price sensitivity (D'Aveni 2007). The objective of the modeling and presentations in this study is to develop a deep understanding of customers' tastes and buying behavior. Such

understanding assists businesses in reaching various groups, using the differences to guide different strategies for increasing brand engagement for each group (Fan & Gordon 2014). Since the success of these strategies depends on the value that customers place on features (D'Aveni 2007), the authors designed *ReviewMap* to develop a list of the features offered by the different brands in the market, and then gather data on how customers perceive those features utilizing the probabilistic topic modeling technique, Latent Dirichlet Allocation (LDA) (Blei, 2012; Bendle, N.T., and Wang, X. 2016).

Marchionini (2006) describes these objectives as searching to learn. Learning searches involve multiple iterations and return sets of objects that require cognitive processing and interpretation. These objects may be instantiated in various media (graphs, or maps, texts, videos) and often require the information seeker to spend time scanning/viewing, comparing, and making qualitative judgments.

With these research-based guidelines in mind, the next section focuses on identifying the text mining and data visualization tools that best allow the generation of feature lists and customer perceptions by interacting with an archive of online reviews.

2. METHOD

TripAdvisor (www.tripadvisor.com) is the world's largest travel site. Its archive contains over 150 million reviews of hotels, restaurants and attractions, many with traveler-supplied photos. Wang (2010) presents a data set of 235,793 hotel reviews collected from *TripAdvisor* in late winter 2009. The data set contains both the free-form text and the 1-to-5 numeric ratings for the following attributes: value, room, location, cleanliness, check in/front desk, service, business service, and overall.

This study used a sample set of reviews for three hotels located in midtown Manhattan, with the pseudonyms "Belluno", "Firenze" and "Venezia". At the time of this writing, these hotels rank between the 17th and 21st percentiles out of 452 New York City hotels. The Firenze and Venezia are head-to-head competitors for the high-end visitor, while the Belluno caters to a more cost-conscious clientele. At the time of this writing, a single, mid-week night, with 60-day advance booking, costs \$365, \$579, and \$629, respectively. These hotels were chosen in order to test the reviews' ability to discriminate amongst closely-ranked players in both the same and different sectors of the competitive landscape.

If online reviews are to serve as a source of competitive intelligence, then the reviews must provide data that discriminates amongst market entities and leads to actionable conclusions. Essentially, the value of the reviews depends upon how well one can predict a market entity from its reviews. In machine learning terms, the reviews need to serve as inputs to a classification model with market entity as the target variable.

Classification models generate predictions by processing nominal and numeric inputs through algorithms such as the neural network or the decision tree. The first task, then, is to convert the review text into vectors of suitable inputs. The most commonly used

method of conversion is to parse the review collection into a term-document matrix A , where a_{ij} = the weight of term i in review j . Weighting schemes ranging from term frequency through log-entropy are available (Berry & Browne, 2005).

Pre-processing was done in R (cran.r-project.org) using the text mining packages *tm*, *slam*, *RWeka*, *NLP* and *openNLP*. LDA analysis was done using the MALLET machine learning for language toolkit (MALLET 2002). Random forest classifiers were built using the R package *randomForest*, and feature/opinion pairs were generated by the authors' own R-implementation of the Garcia-Moya, Anaya-Sanchez & Berlanga-Llavori (2013) model. For information on random forests, see Hastie, Tibshirani, & Friedman (2011) and Breiman & Cutler (2002). Pre-processing consisted of removal of common words such as "the", "at", "also", etc., and hotel-specific words such as "Belluno", "Firenze," and "Venezia". All letters were made lowercase, and trailing spaces were added after punctuation marks that did not have them.

2.1. Summarizing The Term-Document Matrix

Term-document matrices are always high-dimensional and sparse, making them difficult to use as input to a classification model. In practice, the term-document matrix is summarized before processing by the classifier. The summarization utilized in this paper was Latent Dirichlet Allocation (LDA), one of the simpler algorithms for probabilistic topic modeling (Blei, Ng, & Jordan 2003). The following description of LDA is suitable for the purposes of this study.

LDA decomposes a collection of documents into topics -- probability distributions over terms --and represents each document with a (weighted) subset of topics. When fit to a set of documents, the topics are interpretable as themes in the collection, and the document representations indicate which themes each document is about. Thus, the learned topics summarize the collection, and the document representations organize the corpus into overlapping groups. (Chaney & Blei 2012)

So, LDA transforms each review into a vector of weights showing the strength of each topic in the review, where a topic is a probability distribution over the set of terms used in all reviews in the archive. An LDA topic is usually represented by showing a handful of the highest probability terms in the distribution. For example, the terms *room*, *comfortable*, *beds*, *large*, *clean*, *king* might come together in an LDA topic, and the extent to which these terms are used in a given review will be reflected in the size of the weight corresponding to the topic in the vector representing the review.

Machine learning classification models are part of what Breiman (2001a) calls the *algorithmic modeling culture*. The approach is to find an algorithm that operates on inputs to predict responses. The approach was developed to work on complex prediction problems, such as speech recognition and computer vision, where it was obvious that traditional data models such as linear regression were not applicable.

Like least-squares linear regression, algorithmic classifiers are fit to a data set, called the "training set", consisting of both inputs and responses. Rather than goodness-of-fit, the fit is guided by predictive accuracy, and tested on hold-out cases, called the "test

set”, that were not considered in fitting the parameters of the algorithm. Dozens of algorithms have been developed over the past three decades, and model selection is still an active area of research (Shalizi, 2014).

For market segmentation, the authors seek an algorithm that provides both accuracy and interpretability, two qualities that often conflict. Without interpretability, the competitive factors that matter to consumers cannot be identified, and without accuracy, one runs the risk of choosing irrelevant factors upon which to compete or of overlooking important factors completely. One algorithm that provides an effective compromise between accuracy and interpretability is the random forest (Breiman, Friedman, Olsen, & Stone 1984; Breiman 2001b).

2.2. Inclusion of Features and Opinions

In order for a manager to respond to the results of an automated classifier, detail about the review contents is required. Some authors propose linking LDA topics to the individual documents (Chaney & Blei 2012). This study proposes to also include an additional level of information between the LDA topics and the reviews that relies upon concept-level sentiment analysis (Cambria 2014).

Unless complaining about a specific incident, reviewers offer multiple opinions about multiple features of a product or service. These feature/opinion pairs are essentially the review contents of interest for market segmentation. As a result, the authors propose to link an LDA topic to the feature/opinion pairs contained in the high-scoring reviews for the topic. For context, links are provided to the sentences that contain the feature/opinion along with the full review text.

Garcia-Moya, Anaya-Sanchez, & Berlanga-Llavori (2013) present a novel methodology for retrieving product features from a collection of free-text reviews. The method relies upon a language modeling framework that combines a probabilistic model of opinion words and a stochastic mapping model between words to approximate a language model of features. The model was used to generate feature/opinion pairs and measure the strength of the pair in the review using the authors’ own scoring function based upon the product of the feature word’s probability times the sum of the probabilities of the opinion words matching the feature.

2.3. Visual Analytics

A market segmentation based upon the text mining tools and procedure previously discussed must process voluminous data from multiple sources. Such analyses are often done through visual analytics, a collection of tools and techniques that combines machine analysis with the human ability to perceive patterns and draw conclusions (Fan & Gordon 2014). The combination of human and machine strengths supports synthesis, exploration, discovery, and confirmation of insight from data.

A commonly used interface design for visual analytics is the dashboard, where multiple displays are used interactively to interrogate the underlying data. In *ReviewMap*, the integrated displays are the LDA topics, feature/opinion pairs, relevant sentences, and

full review text. In Schneiderman (1996), a visual information seeking mantra of *overview first, zoom and filter, then details on demand*, is presented. For overview of the multi-dimensional LDA topics, the parallel coordinates plot is described as a *clever innovation that makes some tasks easier, but takes practice for users to comprehend*. The authors supplemented the parallel coordinates plot in the dashboard with a plot of a two-dimensional projection of the full-dimension LDA topic data set. The GGobi software package (Cook & Swayne 2007) performs an interactive grand tour (Cook, Buja, Lee & Wickham 2008) of a high-dimensional space from which a user can choose a particular projection with good separation of points.

3. RESULTS

The dashboard is presented using *Tableau* (www.tableausoftware.com), an interactive graphics platform suitable for Schneiderman's *zoom/filter and details on demand* functions. A list of feature/opinion pairs with scores was added, and bar charts showing numbers of reviews and feature/opinion pairs under consideration to the parallel coordinates plot and two-dimensional projection plot of the LDA topics. Zoom/filter (selection) specifications are done within one on-screen panel, with results then propagated to the other panels. Drilling down to relevant sentences and full review text is also done on-screen.

Figure 1: The *ReviewMap* Dashboard

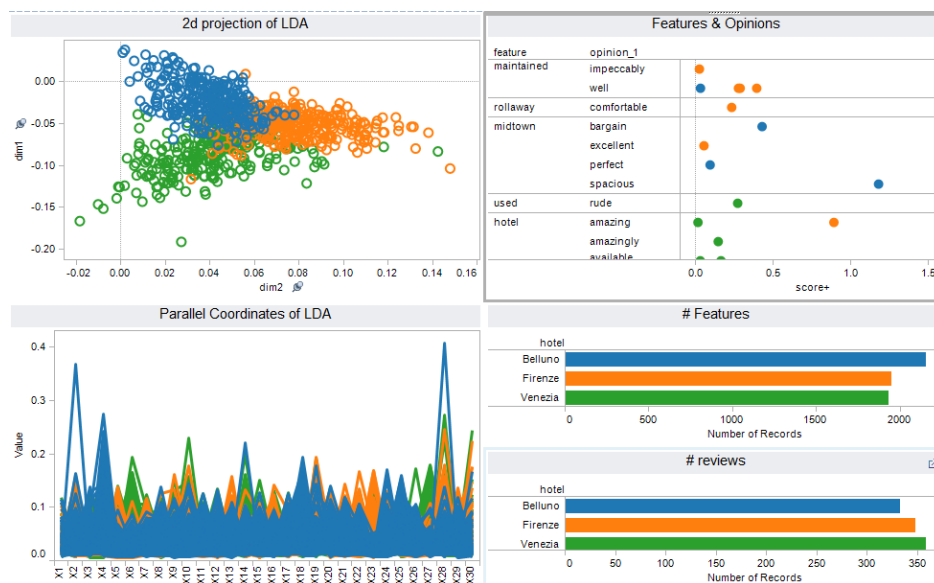


Figure 1 shows a screen capture of the dashboard. The underlying sample data set contained 1,040 free-form text and numeric ratings, which were distributed among the three hotels as follows: Belluno (n=333), Firenze (n=348), and Venezia (n=359). The number of LDA topics was 30 (X1-X30), with representations of LDA analysis appear

on the left. (See Appendix A for a complete listing of LDA topics with associated terms.) The “2d projection of LDA” (upper-left) depicts the clustering of the three hotels between various combinations of tested dimensions, with Belluno tending toward the upper-left, Venezia the lower-left, and Firenze to the right. The “Parallel Coordinates of LDA” (lower-left) shows the values across each of the 30 LDA topics. The “Features & Opinions” panel (upper-right) shows sentiment analysis results, with and the lower-right shows the number of features and reviews among the three hotels currently reflected in the other panels.

The following typical SWOT Analysis questions are considered in the following sections:

- How well can a hotel be predicted based on its reviews?
- What is a given hotel’s greatest perceived strength and weakness? How important is that feature to competitors’ customers?
- Which features are important to all reviewers, regardless of hotel?
- Which features are common to customers of the two most directly competing hotels? Which features are unique? And can competitive actions be uncovered through an analysis of these features?

To illustrate the methods and nature of the results, these questions will be answered for a few of the features uncovered by the text mining. A full SWOT Analysis would involve answering these questions for all discovered features. A full list of these features (with actual spellings) is included in the appendix. Some groups of words are easily interpretable, while others are more obscure. Users of *ReviewMap* need to experiment with the number of features used to find a useful number of features for their particular collections of reviews.

3.1. Predicting a Hotel Based On Reviews

The first test of the discriminating power of the review text is a comparison of classifier performance with various input data sets and the hotel as target. The following was considered:

- A. Naive guessing (i.e., always choose Venezia because it has the highest frequency),
- B. Random forest applied to the “1-to-5” numeric ratings,
- C. Random forest applied separately to a 30-dimensional LDA analysis,
- D. Random forest applied to the LDA and numeric rating variables combined,
- E. Random forest applied to the feature/opinion pairs alone, and
- F. Random forest applied to the LDA and feature/opinion pairs combined

The results are summarized in Table 1 below. The reviews for these three hotels were overwhelmingly positive, as approximately 90% of the numeric Overall Ratings were “4’s” or “5’s” out of five for each hotel.

Table 1: **Random Forest Results For Various Input Data Sets**

Method	Accuracy	Error Rate
Naive guess	34.5%	65.5%
Random forest applied to 1-5 numeric ratings	42.4%	57.6%
Random forest applied separately to 30-dimensional LDA	78.4%	21.6%
Random forest applied to LDA & numeric ratings combined	79.3%	20.7%
Random forest applied to features/opinion pairs alone	48.8%	51.2%
Random forest applied to LDA & features/opinions combined	77.5%	22.5%

From these results, the authors concluded the following:

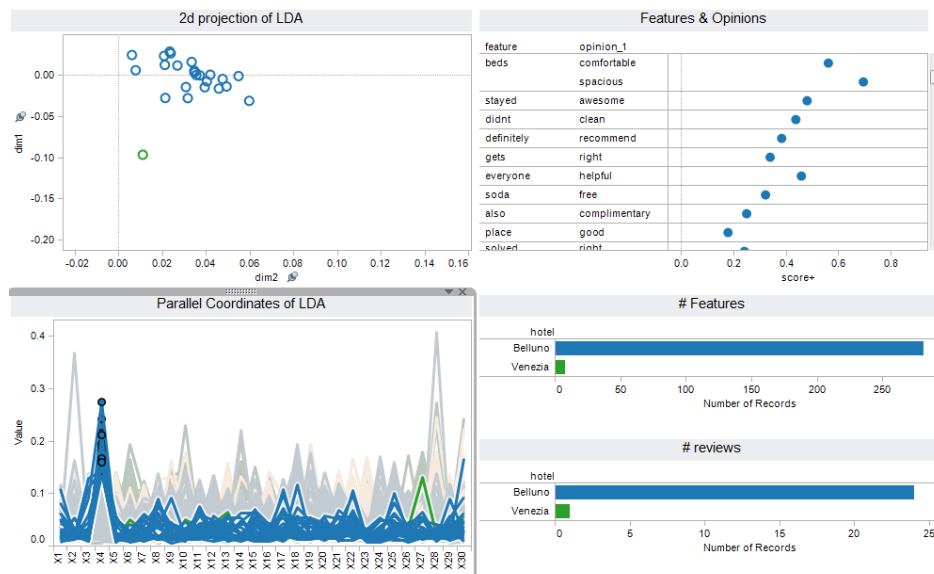
- Reviews carry information specific to each hotel since the accuracy rate increase substantially when text mining was applied.
- Free-form text carries the bulk of the information because the accuracy rate increased substantially when predictive modeling was applied to the text-based factors.
- LDA topics are the best path into the archive since adding features and opinions did not generate any increase in predictive accuracy.

3.2. Identifying Strengths, Weaknesses, and Their Relative Importance

Our method of mining the online reviews to support SWOT Analysis and market segmentation involved (1) drawing up a list of the features offered by all three hotels, and then (2) gathering data on how customers perceive those features among the hotels. First, the features of interest to one hotel’s customers, but not the others, were considered. This can be done by examining the high values of each LDA topic for homogeneous content. This is done in Tableau through the use of an inclusion filter. Figure 2 shows the filtered results on the highest scoring reviews on LDA Topic 4: *breakfast, coffee, free, tea, fruit, eggs...* Clearly, Belluno dominates the topic, and a review of the corresponding feature/opinion pairs and matching sentences reveals further details. Their customers appreciated the free availability of a sit-down breakfast, and an “on the run” breakfast bag. They also noted the variety of items available, along with 24-hour free coffee and tea in the lobby, free Internet, business center and fitness center. Drilling down to the free-form text revealed a collection of comments similar to the following:

“This hotel boasts a free breakfast which runs from 6am-10am. There are two dining rooms full of tables and chairs and the decor (here and throughout the hotel) is a very nice contemporary; comfortable design. The hotel changed their hot breakfast selections daily (waffles/sausage; eggs/sausage; etc.). There was also a variety of bakery items such as biscuits; cinnamon biscuits; sweet bread; muffins; bagels and toast. Other items offered were: ready to make oatmeal; cereal; fresh fruit and fruit salad as well as a variety of juices; milk; coffee and hot chocolate. It would be difficult for someone to not find something they like!”

Figure 2: **High Scoring Reviews on LDA Topic 4 (breakfast, coffee, free, tea, fruit, eggs)**



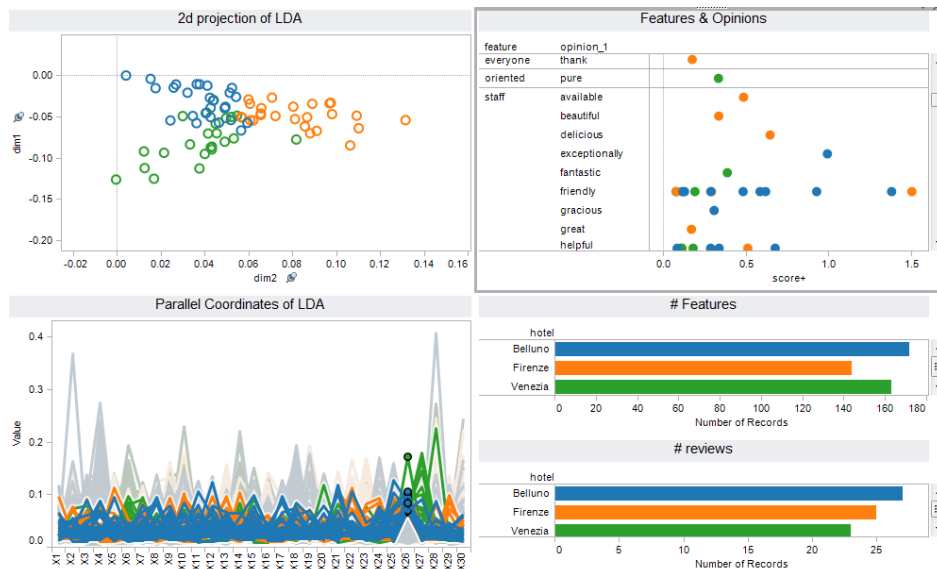
This information can be used for competitive actions through advertising and staff development. The Belluno website displays information about free amenities on its home page. Every piece of communication to customers should do the same, and the importance of these amenities should be included in all staff training programs. These results should also make their way to top management for firm-wide resource allocation decisions. A decision-maker considering the elimination of some, or all of these, free amenities in an effort to reduce costs should be aware of just how important the amenities are to customers, and how closely customers identify the amenities with the Belluno brand.

Apparently, some hotels are missing this relatively easy-to-find information. A leading LDA topic for Venezia customers is LDA Topic 6: *view amazing fantastic* (specifics of view withheld). These high-scoring reviews refer to the view of a landmark NYC site from many of the rooms. Many reviewers post snapshots of this view on TripAdvisor, but there is no picture of the view on the Venezia website. The enthusiasm shown below is typical of the reviews scoring well on this topic, which the random forest algorithm identified as having the highest power of discrimination for Venezia.

“The rooms are modern and comfortable; but the best feature of every room is the OUTSTANDING view. Even the gym faces the landmark NYC site I would definitely recommend this hotel to anyone. To see for yourself; I posted a video of my stay... When we arrived at the hotel; we were awed by the beautiful lobby and restaurant spaces; especially the main entrance; which faces NYC landmark. The view of the landmark from the gilt lobby area is truly magical (made us feel as if we were in Europe). I had stayed in competing hotel the previous year; and although I found the common spaces there to be lovely; I'd have to say that the Venezia beats out the competing hotel; especially given the view of the landmark.”

LDA topics whose high value reviews have heterogeneous content represent items of interest to all reviewers across the three hotels. Figure 3 shows filtered results on high scoring reviews for Topic 26: *staff, helpful, friendly, stay, wonderful, extremely*. One can see in the “Features & Opinions” panel (upper-right) that Belluno dominates the staff/friendly feature/opinion pair. Does this mean Firenze and Venezia staff are unfriendly? A review of the full list of feature/opinion pairs reveals that Firenze and Venezia reviewers tend to comment on individuals rather than the “staff”. These reviews name check-in, housekeeping, concierge, bellman, etc. when describing friendly and helpful encounters, and the authors conclude that friendly staff are found at all three hotels. Whether the linguistic differences in the reviews are due to differences in the reviewers or staffs requires further investigation. As an aside, it was also determined that *unfriendly* and *unhelpful* opinion pairs occurred with very low frequency and equally across the three hotels.

Figure 3: **Topic 26 (staff, helpful, friendly, stay, wonderful, a topic whose high scoring reviews are heterogeneous)**



3.3. Identifying Features Important to All Reviewers

The analyses presented so far follow a path from linguistics, i.e. the LDA topics, to features/opinions to free-form text to draw conclusions about competitive relationships. The analysis that follows begins with the apparent competitive relationships shown in the two-dimensional projection of the LDA topics and searches the corresponding feature/opinion pairs for explanatory details.

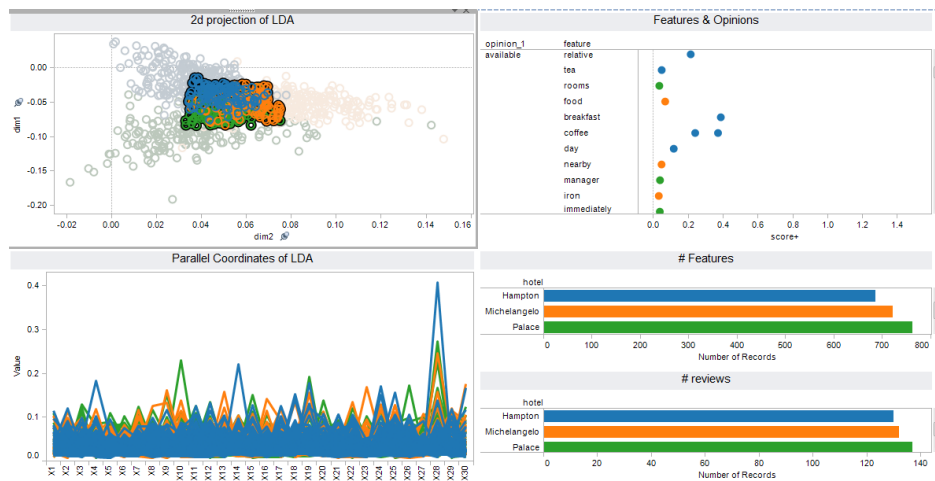
Figure 4 shows a filter of the middle of the 2d projection cloud. The authors looked to these reviews to find factors important to all reviewers regardless of hotel. Comparing the parallel coordinates plot in Figure 4 to that of Figure 1, one can see each hotel's

“true believers” have been filtered out. It is noted that a Firenze review shows at the top of Topic X4 (breakfast, coffee, free, tea, fruit, eggs) in Figure 4, where in Figure 1, that topic is dominated by Belluno. Heterogeneity can now also be seen in Topic X6 (view, amazing, fantastic), originally dominated by Venezia, and in Topic X23 (beautiful, suite, upgraded, special), originally dominated by Firenze.

An examination, therefore, of the feature/opinion pairs for these filtered reviews should reveal the topics necessary to compete at the 20th percentile of New York City hotels. Value for the money, room cleanliness, bed, bath, service, and location were cited by all in mostly positive comments. Negative comments were posted by those who found these features lacking.

In this center cloud of points, it was not possible to distinguish between higher-priced Firenze and Venezia reviewers. It was, however, still possible to identify many Belluno reviewers through their enthusiasm for the free amenities and for the hotel’s location near the subway (LDA Topic X18).

Figure 4: A Subset of The Most Similar Reviews That Span All Three Hotels



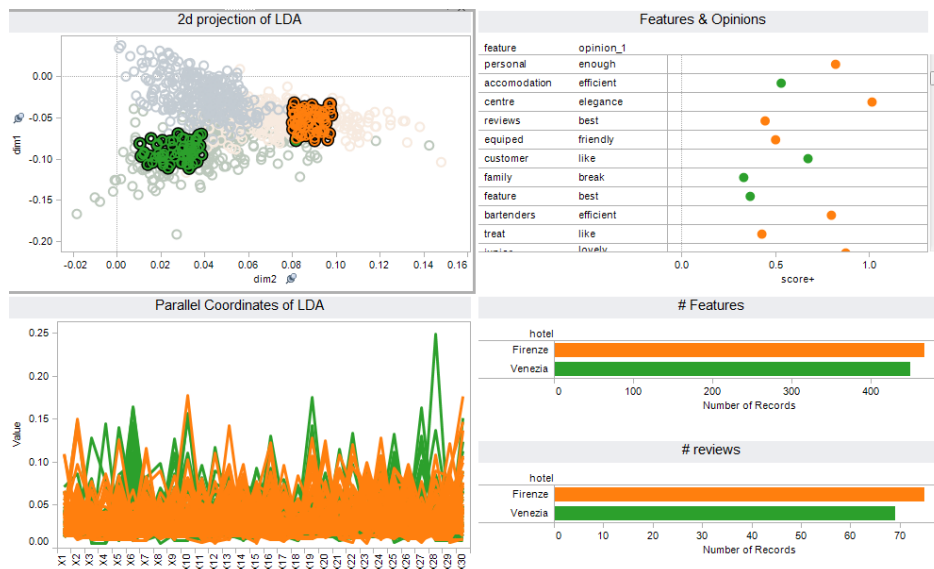
3.4. Comparing Significant Competitors

The next analysis focuses on the Firenze vs. Venezia competition by selecting a subset of reviews from the heart of each hotel’s cloud of points in the “2D projection of LDA” panel, as shown in Figure 5. These two hotels were selected due to their similar pricing, relative to the three hotels. In order to see what was being mentioned in the 145 reviews selected, the authors drilled down to the features extracted, which revealed 10 features that were mentioned 10 times or more. They were, in alphabetical order: “Bathroom”, “Bed”, “Breakfast”, “Hotel”, “Location”, “Room”, “Service”, “Shopping”, “Staff”, and “Stay”. To assess how reviewers perceive these features, the authors then drilled down further to the sentences containing these features in *Tableau*.

For the feature *breakfast*, reviewers were equally negative regarding both hotels, complaining about cold food, crowded conditions, high cost of optional items, such as orange juice, and slow service. For many, the breakfast was the only negative feature in the review. Clearly, the Firenze and Venezia reviewers either have different standards or different experiences than the Belluno reviewers who thought the free breakfast was one of that hotel's highlights.

For the feature *room*, reviews of standard rooms were fair. Additional services such as turn-down, morning newspaper, and flowers or chocolate upon arrival improved the tone of the review, and any kind of upgrade generated genuine enthusiasm. It was not possible to determine whether these upgrades occurred according to a hotel policy or at the discretion of the staff person, but based upon this feedback, an upgrade has significant impact on both Firenze and Venezia customers. Further investigation in how best to exploit this feature is indicated.

Figure 5: A Query of Mainstream Venezia and Firenze Reviews



For Firenze vs. Venezia reviews, the original feature-opinion sort order was reversed in *Tableau* to opinion-feature. This allowed the examination of the distribution of hotel within feature/opinion pairs that contained superlatives such as *amazing*, *elegant* and *impeccable*. While the frequency of superlatives was similar for both hotels, a clear difference was found in the type of superlatives used. Firenze reviewers led in the use of *beautiful*, *elegant*, *gorgeous*, and *luxurious*, while the Venezia reviewers led in the use of *amazing*, *fabulous*, and *fantastic*. An examination of the features paired with these superlative opinion words showed that Firenze reviewers used superlatives to talk about hotel features such as the *room*, *bed* and *bathroom*, while the Venezia reviewers used superlatives for less intimate aspects of the hotel such as the *view*, *shopping* and *location*. A drill-down on the *room*, *bed* and *bathroom* features for Venezia reviewers

showed mostly satisfactory reports, some complaints, and almost none of the superlatives used by the Firenze group.

Firenze and Venezia management may have actionable intelligence in this observation. If the actual reason for the observation is more luxurious interiors at the Firenze, then this in addition to their lower price and equivalent location could provide the basis for differentiating the Firenze from the Venezia. Other explanations of the observation are certainly possible, but Venezia management may want to investigate whether this difference corresponds to a genuine perception of fading interior quality amongst its customers.

4. DISCUSSION

Through the SWOT Analysis and customer segmentation of 1,040 *TripAdvisor* reviews for three hotels located in midtown Manhattan, the authors conclude that online reviews can provide managers with new, previously unknown, but yet competitively valuable, insight and feedback. Since online review archives are too large to process manually, the value that online reviews deliver to managers is a function of the ability of text mining tools to support realistic analyses.

Through the SWOT framework, hotel management can analyze customer perspective of their operations, as well as their competition. Actionable intelligence can be developed regarding areas of operation that may require improvement. In addition, hotel-related features that are valued by customers can be identified that may differentiate a given hotel from the competition, after which, appropriate strategic and/or operational actions can then be taken.

4.1. Text mining tools to support realistic analysis

Tukey (1977) can be looked to for a realistic assessment of what *ReviewMap*, given its current state, brings to management practice. *ReviewMap* utilizes modern computer technology to look at data to see what it seems to say, assisting the user in recognizing partial descriptions, and then looking beneath them for new insights. The system processes an archive of online reviews spanning multiple suppliers into a hierarchy of data of increasing dimensionality in order to make it more easily and effectively understandable. The levels of hierarchy, in increasing order of dimensionality, are: *competitors*, *topics*, *features/opinions*, *feature-relevant sentences*, and *full review text*. Previous research, including Chaney & Blei (2012), correctly recognized the need for such a hierarchy, which provides a platform that connected topics to full text. While this hierarchy provides valuable assistance to one browsing a collection, the authors found it too shallow to gather competitive intelligence within a reasonable time-frame.

The inclusion of the two-dimensional projection of LDA topics, and the feature/opinion pairs and their corresponding sentences allows one to ask questions, such as, "What is our greatest strength, and how important is that feature to our competitors' customers?" The search for an answer begins with an LDA topic on which customers score highly. The collection of all reviews that score highly on this topic may be homogeneous or

heterogeneous. Heterogeneous is a simpler case because customers and competitors' customers are talking about the same features, and the authors can identify the marks of each in the feature/opinion list by the color coding. If the high scoring topics are homogeneous, then it is necessary to check that competitors' customers are not using different language to address the same features of the topic prior to concluding the feature is of interest to *only* customers of a particular organization.

ReviewMap results such as these demonstrate the capabilities of text mining tools to support realistic analyses. The authors were able to achieve classifier accuracy of almost 80% from the review text on a relatively small, overwhelmingly positive, data set. In addition, *ReviewMap* uncovered exploratory evidence of perceived strengths and weaknesses of each hotel, and provided intelligence for the consideration of competitive actions in advertising, staff development, customer service, brand quality, and physical appearance.

The analysis and modeling presented in this paper illustrates that the proper use of natural language processing models requires a combination of machine and human intelligence when drawing conclusions. Further, a visualization platform utilizing a point-and-click interface with flexible query and rich display capabilities is essential. Organizations using *ReviewMap* and similar text mining/visualization systems must manage their human resource commitment to the effort. Suggestions for this include direct, hands-on use of *ReviewMap* by domain experts rather than IT staff, as the more people in the chain from pre-processed data to management action, the greater the chance of misinterpretation, omission, and misplaced emphasis.

In addition, the authors found that doing the analyses using pre-defined questions and search procedures was more productive than bottom-up browsing. However, it is acknowledged that this observation could be a result of the authors' own cognitive styles, and may not apply to all users. Effective use of the tool, therefore, must include training and practice to develop each user's style of use, as one size definitely will not fit all. Lastly, an evaluation of how "realistic" any analysis is must include the amount of time and effort required to conduct the analysis. This is an important consideration, as excessive demand on the analyst could conceivably deem such work impractical. The most noticeable demand was the amount of time required to understand the contents of the review collection, which is not surprising. The Searching to Learn tasks described in Marchionini (2006) are clearly time consuming. The CRISP-DM data mining methodology includes Data Understanding as a major early step in any data mining project (Shearer 2000). Further, Breiman (2001a) advises us to "live with the data before plunging into modeling."

4.2. Further Research

As previously indicated, working with *ReviewMap* is an exploratory activity. As such, confirmatory analyses must follow relatively new procedures that attempt to adapt the concepts of test statistics, tests, null distributions, significance levels and p-values to visualization results (Buja et al 2009; Hubbard 2014). These procedures are beyond the scope of this paper, and will be considered in future work.

The next step after a *ReviewMap* analysis could be a simple action such as the addition of a few images on a company's website or something more complex such as a pilot project to test new upgrade policies. If the user fears the authors' multiple comparisons may be yielding spurious associations, confirmatory study can be conducted through monitoring future reviews for consistency with current observations, or by using surveys and other data collection methods. Traditional statistical inference techniques, or emerging inference methods that rely upon visual data (Buja 2009), are applicable depending upon the data collected. Another option, available after confirmation studies, is to include *ReviewMap* observations in a Bayesian decision analysis to update models of uncertainty for decision-making metric values (Hubbard 2014).

REFERENCES

- Bendle, N.T. and Wang, X. (2016), "Uncovering The Message From The Mess of Big Data", *Business Horizons*, Vol. 59, No. 1, pp. 115-124. DOI: <http://dx.doi.org/10.1016/j.bushor.2015.10.001>
- Berry, M.W. and Browne, M. (2005), *Understanding Search Engines: Mathematical Modeling and Text Retrieval* (2nd ed.), Society for Industrial and Applied Mathematics, Philadelphia, PA. DOI: <http://dx.doi.org/10.1137/1.9780898718164>
- Blei, D.M. (2012), "Probabilistic Topic Models", *Communications of the ACM*, Vol. 55, No. 4, pp. 77-84. DOI: <http://dx.doi.org/10.1145/2133806.2133826>
- Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.
- Breiman, L. (2001a), "Statistical Modeling: The Two Cultures", *Statistical Science*, Vol. 16, No. 3, pp. 199-231.
- Breiman, L. (2001b, January), *Random Forests*, University of California, Berkeley, CA.
- Breiman, L. and Cutler, A. (2002), "Random Forests - Classification Description", viewed August 16, 2014, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
- Breiman, L., Friedman, J.H., Olsen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.
- Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. and Wickham, H. (2009), "Statistical Inference For Exploratory Data Analysis and Model Diagnostics", *Philosophical Transactions of The Royal Society A*, Vol. 367, pp. 4361-4383. DOI: <http://dx.doi.org/10.1098/rsta.2009.0120>
- Cambria, E. (2014, June 1), Concept-Level Sentiment Analysis, viewed August 19, 2014, <http://sentic.net/tutorial/>
- Chaney, A.J. and Blei, D.M. (2012), Visualizing Topic Models, *Proceedings of the International AAAI Conference on Social Media and Weblogs*.
- Cook, D., Buja, A., Lee, E.-K., & Wickham, H. (2008), Grand Tours, Projection Pursuit Guided Tours and Manual Controls. In C.-H. Chen, W. Hardle, & A. Unwin (Eds.), *Handbook of data visualization* (pp. 295-314). Heidelberg and Berlin, Germany: Springer. DOI: http://dx.doi.org/10.1007/978-3-540-33037-0_13
- Cook, D., and Swayne, D. (2007), *The GGobi book*, Springer, New York, NY.
- D'Aveni, R.A. (2007, November), "Mapping Your Competitive Position", *Harvard Business Review*, pp. 110-120.
- Fan, W. and Gordon, M.D. (2014), "The Power of Social Media Analytics", *Communications of the ACM*, Vol. 57, No. 6, pp. 74-81. DOI: <http://dx.doi.org/10.1145/2602574>
- Garcia-Moya, L., Anaya-Sanchez, H. and Berlanga-Llavori, R. (2013), Retrieving Product Features and Opinions From Customer Reviews. *IEEE Intelligent Systems*, Vol. 28, No. 3, pp. 19-27. DOI: <http://dx.doi.org/10.1109/MIS.2013.37>
- Hastie, T., Tibshirani, R. and Friedman, J. (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer, New York, NY. DOI: <http://dx.doi.org/10.1007/b94608>
- Hubbard, D.W. (2014), *How To Measure Anything: Finding The Value of Intangibles In Business* (3rd ed.), Wiley, New York, NY.

- Ittoo, A., Nguyen, L.M. and van den Bosch, A. (2016), Text Analytics In Industry: Challenges, Desiderata and Trends. *Natural Language Processing and Text Analytics in Industry*, Computers in Industry, Vol. 78, pp. 96-107. DOI: <http://dx.doi.org/10.1016/j.compind.2015.12.001>
- Köseoglu, M.A., Ross, G. and Okumus, F. (2016), “Competitive Intelligence Practices In Hotels”, *International Journal of Hospitality Management*, Vol. 53, pp. 161-172.
- Marchionini, G. (2006), “Exploratory Search: From Finding To Understanding”, *Communications of the ACM*, Vol. 49, No. 4, pp. 41-46. DOI: <http://dx.doi.org/10.1145/1121949.1121979>
- McCallum, A. (2002), MALLET homepage, viewed August 22, 2014, <http://mallet.cs.umass.edu/>
- McFarland, L.A. and Ployhart, R.E. (2015), “Social Media: A Contextual Framework to Guide Research and Practice”, *Journal of Applied Psychology American Psychological Association*, Vol. 100, No. 6, pp. 1653-1677. DOI: <http://dx.doi.org/10.1037/a0039244>
- Quinlan, J.R. (1986), “Induction of Decision Trees”, *Machine Learning*, Vol. 1, No. 1, pp. 81-106. DOI: <http://dx.doi.org/10.1007/BF00116251>
- Shalizi, C.R. (2014, June 11), Model Selection, viewed August 16, 2014, <http://vserver1.cscs.lsa.umich.edu/~crshalizi/notabene/model-selection.html>
- Shearer, C. (2000), “The CRISP-DM Model: The New Blueprint For Data Mining”, *Journal of Data Warehousing*, Vol. 5, No. 4, pp. 13-22.
- Shneiderman, B. (1996), “The Eyes Have It: A Task By Data Type Taxonomy For Information Visualizations”, *Proceedings IEEE Symposium on Visual Languages*, pp. 336-342. DOI: <http://dx.doi.org/10.1109/VL.1996.545307>
- Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.
- Wang, H. (2010), “Latent Aspect Rating Analysis On Review Text Data: A Rating Regression Approach”, *The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '2010)*, pp. 783-792. DOI: <http://dx.doi.org/10.1145/1835804.1835903>
- Zeng, D., Chen, H., Lusch, R. and Li, L.-I. (2010), “Social Media Analytics and Intelligence”, *IEEE Intelligent Systems*, Vol. 25, No. 6, pp. 13-16. DOI: <http://dx.doi.org/10.1109/MIS.2010.151>

APPENDIX

LDA Topic	High Scoring Terms
1	day stay city staying visit long back corner mid making start decided disappointed year visited offer forward typical son
2	great place time nice don show good convenient back town ll awesome italian shopping fun husband problems deli fine
3	desk front hotel time busy full guests housekeeping line staff person provided areas area issue addition job choices regular
4	breakfast coffee hot free tea included fruit fresh pastries complimentary water cookies chocolate cold selection eggs buffet apples items
5	service restaurant concierge excellent weekend bar der expected table shopping drink corner doormen recommended reservations professional standards pricey makes
6	view st fantastic cathedral floor amazing patrick expensive overlooking lovely window fabulous views patricks athing absolutely side grand superb
7	hotel rooms perfect huge extremely luxurious appointed prices awhere beat bathrooms beautifully half visiting croissants immaculate word recommend mother
8	good hotel location clean size week choice expect nearby decent happy ideal attractions july average sit min definately summer
9	stayed hotel nights found wife years door standard end polite previous class high complimentary fault lobby april experience october
10	hotels quality business rooms midtown ve find top luxury modern decor property feel rates fine rate style hilton amenities
11	great location staff clean big spacious friendly comfy incredible super breakfasts choose mornings exceptionally bit keeping deep bathrobes lack
12	room stayed booked spacious problem weekend courteous world site surprised smoking thought renovated pleasantly tired expectations fairly moved furniture
13	nice stay lobby loved enjoyed bar home touch ve part pretty treat favorite mentioned terrific big friend pleased toiletries

LDA Topic	High Scoring Terms
14	hotel made year arrived door daughter travel american return reservation immediately car leave package arrival brought chose offered returned
15	hotel recommend staff family highly aone kids feel places days children holiday tourist sightseeing round vacation families returning gem
16	room service back experience spent building days served charge entire work impressed point city wouldn personal order welcoming bigger
17	breakfast price free small rate internet paid cost good pay crowded gave continental higher heart standards juice area minor
18	subway block street manhattan blocks close times ma place walked shows north easy tour bus station theaters couple safe
19	room check early minutes ready checked arrived pm returned wait call late requested called hours hour booked waiting wasn
20	time people didn morning trip bags night parking city quiet luggage quick cab eat felt taxi ice lots ride
21	location breakfast walking comfortable distance district theater ave theatre restaurants complimentary included excellent quiet plenty worked check adequate house
22	room comfortable beds large bed located clean king towels queen pillows couldn plenty linens sized changed double warm enjoy
23	beautiful suite husband upgraded special wanted perfect birthday looked trip lovely love upgrade kind towers chocolates music radio tower
24	reviews ma lot star thing hotels things bad book deal read give trip based recently tripadvisor due hard hear
25	square times walk central park avenue broadway center rockefeller close short restaurants easy location minute theatres middle pleased met
26	staff helpful friendly stay wonderful extremely needed pleasant found incredibly couple real plentiful attentive departure accomodating upscale trips work
27	food day worth executive evening money access lounge return extra drinks level excellent eat lunch floors ate west expect
28	night told left make asked find manager experience put athing reception called card note treated paying guest maid completely
29	night room noise floor lobby street don high quiet noisy side nicely light wall facing set mind elevators decorated
30	bathroom room area bit tv large bed shower huge air turn bath marble tub screen sleep flat lots space

William J. Amadio, PhD, Associate Professor
Rider University
College of Business Administration
2083 Lawrenceville Road
Lawrenceville, NJ 08648 USA
Telephone: (609) 896-5254
Fax: (609) 896-5304
E-mail: amadio@rider.edu

J. Drew Procaccino, PhD, Associate Professor
Rider University
College of Business Administration
2083 Lawrenceville Road
Lawrenceville, NJ 08648 USA
Telephone: (609) 896-5259
Fax: (609) 896-5304
E-mail: procaccinod@rider.edu

Please cite this article as: Amadio, W.J., Procaccino, J.D. (2016), Competitive analysis of online reviews using exploratory text mining, *Tourism and Hospitality Management*, Vol. 22, No. 2, pp. 193-210, <https://doi.org/10.20867/thm.22.2.3>



Creative Commons Attribution – Non Commercial – Share Alike 4.0 International