# Visualizing Hotel Reviews: a Case Study using TripAdvisor Data

Fabian Colque
Instituto de Informática - UFRGS
e-mail: feczegarra@inf.ufrgs.br

João L. D. Comba
Instituto de Informática - UFRGS
e-mail: comba@inf.ufrgs.br

Viviane Moreira
Instituto de Informática - UFRGS
e-mail: viviane@inf.ufrgs.br

***Abstract*—Finding hotels that are suited to one's needs can be a time-consuming task. In this process, people usually rely on customer reviews from travel websites. These websites typically contain many reviews shown in a textual format and a chart that summarizes the overall opinion about a given hotel. In order to compare a number of hotels, users will need to read many reviews and navigate through many web pages. With the goal of aiding users in this process, in this paper, we propose a visual tool for hotel comparison. The tools focuses on the most important aspects that can be extracted from hotel reviews (location, cleanliness, rooms, etc.) and it allows ordering the hotels by one or more of these aspects. We display aspect information using stacked bar charts alongside their ranking, which becomes very useful for comparing hotels. Additionally, we provide a scatterplot matrix combining aspects to aid in situations in which the users wish to make pairwise comparison of aspects. We developed a web-browser demo of our proposed tool using real data from TripAdvisor and demonstrate how it can be used to perform hotel comparisons in 67 different locations.**

## I. Introduction

Every year, about 150 million hotel bookings are made online[1]. Before committing to a hotel, users typically rely on previous experiences of other users which are expressed in the form of textual reviews. There are some popular websites that contain hotel reviews such as Booking.com, TripAdvisor, Hotel.com, etc. While these websites do a great job in putting together millions of reviews, they still lack user friendly interfaces to enable the comparisons of a number of hotels. For each hotel, they typically show a histogram that allocates the reviews according to their overall rating in a five point scale, and the text of the reviews. If a user wants to compare a number of hotels to make a choice, it is necessary to navigate through several web pages and read many reviews.

The analysis of reviews has gained significant interest in recent years in the areas of sentiment analysis or opinion mining [**?**], [2]–[4], [15]. In its simplest form, the goal is to identify the polarity of the review, i.e., whether it expresses a positive or a negative opinion. A polarity can be attributed to the entire review, to a sentence, or to each *aspect* mentioned in the review. An aspect is an attribute or component of the entity being reviewed. In the hotel domain, for example, the aspects are location, service, rooms, cleanliness, etc. Research on *aspect-based opinion mining* [**?**] aims at extracting, grouping, and determining the sentiment polarities of the aspects mentioned in reviews. Aspects are important in this work as different people may favor a different aspect when choosing a hotel. While some may consider location as the most important factor, others may be more concerned with the services provided by the hotel, or even a combination of these two aspects. Clustering of reviews is described in [7], [8] to find reviews that share similar ideas and how they evolve throughout time. Visualization and interaction techniques are being used to offer insights in the text collection analysis [3]–[5], [9], [10] and can be useful to analyse hotel reviews. Work related to this paper [1], [6] have offered a summarized way to evaluate customer opinions.

In this work, we describe a new approach using visualization techniques to compare hotel reviews. Figure 1 illustrates the main components of the prototype we developed so far. First, the user can select the location using a map or through a pull-down menu. The data for all hotels of a given location are shown using stacked-bar charts to display the information regarding each different aspect, which allows the user to compare hotel results. Also, the visualization allows ordering the data using different aspects, which results in multiple rankings of hotels that are also useful to compare hotels. Finally, a refinement of the hotels selected over a scatterplot matrix of pairwise aspects allows the user to narrow down the analysis into hotels that satisfy a given search criteria.

## II. Data Representation and Visualization Goals

### A. Data

We worked with 235,793 hotel reviews about 12,773 hotels from TripAdvisor[2]. The reviews are further separated in 67 different locations, which in this dataset comprises of different cities across the world. This dataset was selected because it already has the ratings given to the aspects extracted from the reviews – since aspect extraction is outside the scope of this work. Metadata about the hotels include ratings in a scale from 0 to 5 of the following aspects: overall (i.e., the overall opinion about the hotel), value, rooms, location, cleanliness, check in/front desk, service, sleep quality, and business service. Whenever a specific aspect is missing from a given review, the rating is set to -1. There are additional attributes that can be used as part of the visual interaction with

[1]http://www.statisticbrain.com/internet-travel-hotel-booking-statistics/

[2]http://times.cs.uiuc.edu/~wang296/Data/

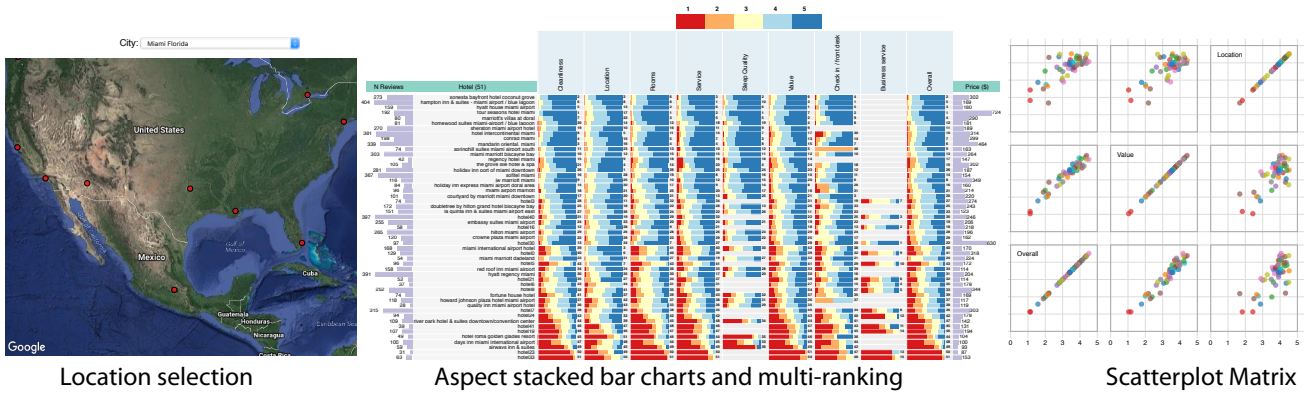


Fig. 1. Main components of the visualization interface: local selection, aspect charts and multiple rankings, and scatterplot matriz.

the user, such as the geographic location of the cities and the number of opinions that a particular hotel has received.

### B. Visualization Goals

The design of our tool was guided by the following of goals, which aim at aiding the user in comparing a number of hotels.

- *Ranking*: it should be possible to rank the hotels according to each aspect and any combination of aspects.
- *Filters*: Users should be able to apply different types of filters. Initially, the user should be able to choose the destination city. Later, scatterplot allows filtering hotels.
- *Interactivity*: The tool should be intuitive and user friendly.

## III. User Interface and Visualization Techniques

In this Section we describe the user interface and visualization techniques employed to analyze the TripAdvisor data.

### A. Location selection

The first level of interaction in the interface is the selection of the location. We provide two ways to perform this selection. The first one uses a Google maps interface to display a world map with red circles indicating locations with data. The user can pan and zoom into the map, and click over the red circle to select the location. Alternatively, we have a pull-down menu that lists all available locations. This last selection is viable since the number of locations is rather small and can be scrolled quickly. For the purposes of the current dataset, these selection alternatives were adequate. We deferred to implement a textual search for locations for larger datasets.

### B. Display of Ratings Associated with Aspects

The dataset comprises 67 different locations, with a varying number of reviews for 9 different aspects. Each review has a sentiment score from 0 to 5. We use a normalized stacked chart to display the information of each different aspect. The area of each of the 5 bars in this chart is normalized by the percentage of reviews in each sentiment class over the total number of reviews. We display each chart with a divergent color scale of 5 different values, ranging from red (most negative), passing

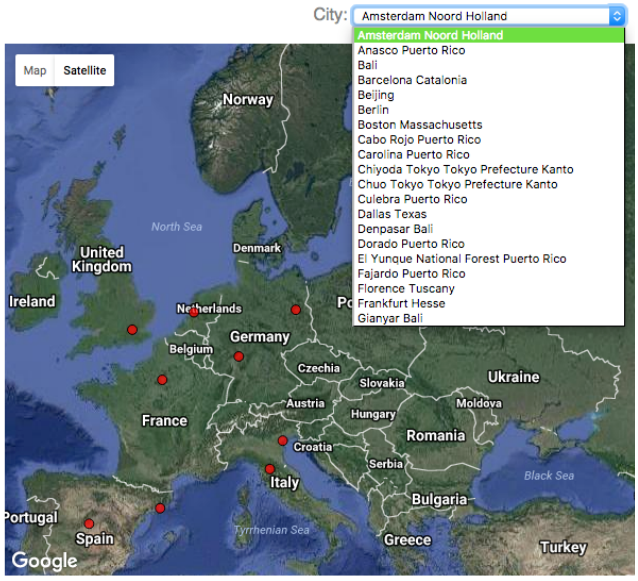


Fig. 2. Location selection. The user can select locations by clicking over red circles in a map, or by scrolling through the list of locations.

through yellow (neutral), to blue (most positive). We display horizontally the charts for the different aspects of a given hotel. Figure 3 illustrates the normalized stacked bar charts we obtain using four different aspects (columns) and 17 hotels (row).

### C. Display and Sorting of Multiple Rankings

One important aspect of the analysis of hotel reviews is the ability to compare hotels based in the results of a given aspect. For example, customers often explore hotels based on price. Therefore, our interface must provide a mechanism to allow the user to sort hotels based on a given aspect. We support this sorting for a single aspect or multiple aspects (selected in a checkbox over each aspect). The ordering using multiple aspects computes the average results of the selected aspects. Currently, we do not support weighted averages, which would allow giving more weight to a given aspect, but such a change could be trivially incorporated in our code. The result is

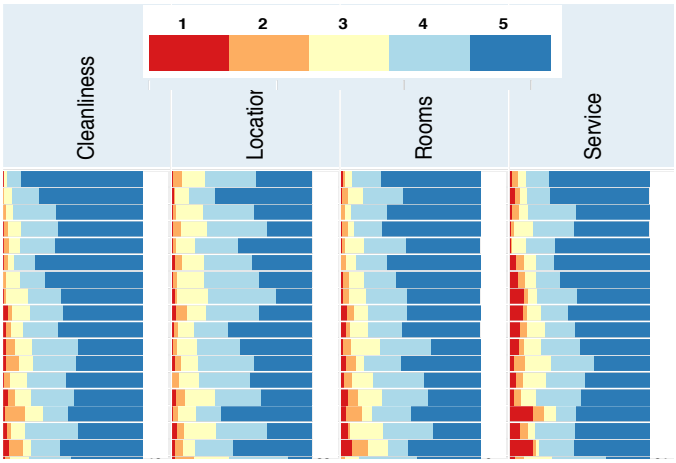


Fig. 3. Display of different aspects using normalized stacked bar charts.

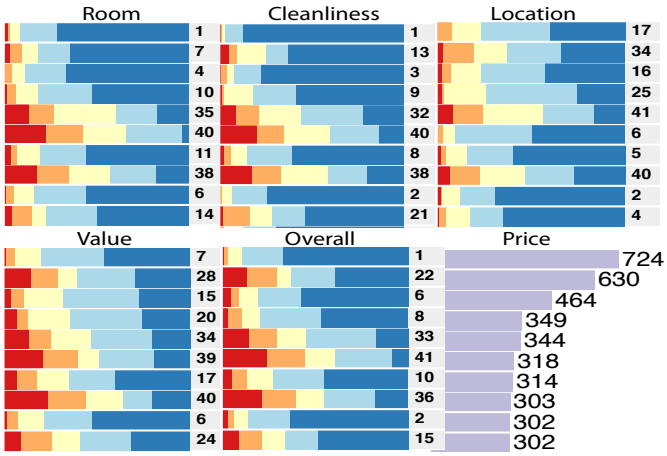


Fig. 4. Multiple rankings and sorting. In this example we show the top 10 hotels in Miami-Florida sorted by price. The individual ranking in each aspect is shown to the right of the chart. Observe that the most expensive hotel is also the first in the ranking for the aspects overall, room, and cleanliness, but it is the $7^{th}$ in the aspect value and $17^{th}$ in the aspect location.

a ranking of hotels based on the chosen criteria. Figure 4 illustrates the top 10 hotels in Miami-Florida sorted by price, from the most to the least expensive.

The sorting of hotels has an impact on the other aspects which is useful in the comparative analysis. For example, the order of a hotel in the ranking by price is not necessarily the same order in the location aspect. In fact, each aspect has its own individual ranking. The problem of displaying multiple rankings is well studied in the visualization community [12]–[14], but there are still challenges on how to display rankings for more intricate data. In our visualization, we display alongside the charts for each aspect, a number that corresponds the ordering of the hotel in the individual aspect ranking. We return to Figure 4 to illustrate the multiple rankings for 5 different aspects. For example, consider the first row of charts in each of the different aspects. They all correspond to the most expensive hotel in Miami, costing $724 dollars. Following the individual rankings alongside each aspect, we observe that this hotel is also the first in the ranking for the aspects room, cleanliness, and overall, but $17^{th}$ in location and $7^{th}$ in value. This multiple ranking view offers an intuitive way for the user to compare the results of each hotel, and consider compromises while choosing a hotel. Looking at the figure, we observe that the $9^{th}$ most expensive hotel is much cheaper ($302 dollars) than the most expensive hotel, while being second in the ranking for cleanliness, location, and overall aspects, and $6^{th}$ in room and value aspects.

### D. Selection using the ScatterPlot Matrix

The comparison of multiple rankings in some situations might display more information that the user needs to make the analysis. This is specially important when the location has a large number of hotels. For example, if the user is concerned with the location and overall aspects, it would be interesting if the analysis could be constrained by hotels that have, for example, the top scores in both of these aspects. To support this additional selection, and make it general to consider multiple aspects, we display the data using a scatterplot matrix,

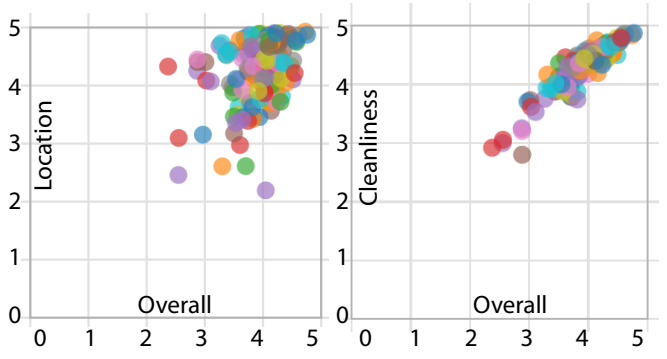


Fig. 5. Two entries in the scatterplot matrix (overall × location and overall × cleanliness). Observe how the overall aspect has a linearly correspondence with the cleanliness aspect.

a construction often used in the visualization community to create pairwise scatterplots of multi-dimensional data.

We display the scatterplot matrix for all aspects of our data, or for a subset of aspects based on user selection. Each entry in this matrix displays a scatterplot for a pair of aspects. The values associated with each aspect correspond to the average of each aspect. The user can directly interact in each cell of the scatterplot matrix by defining a rectangular region of interest. The hotels contained within the selection area are updated in the multiple-ranking visualization.

Figure 5 displays two entries in the scatterplot matrix. While the relation of the overall and location aspects is more distributed, there is a cleat linear correlation between the overall and cleanliness aspects.

## IV. Results

We developed a web-based prototype using D3 [**?**] to validate the concepts proposed in this work. Some examples

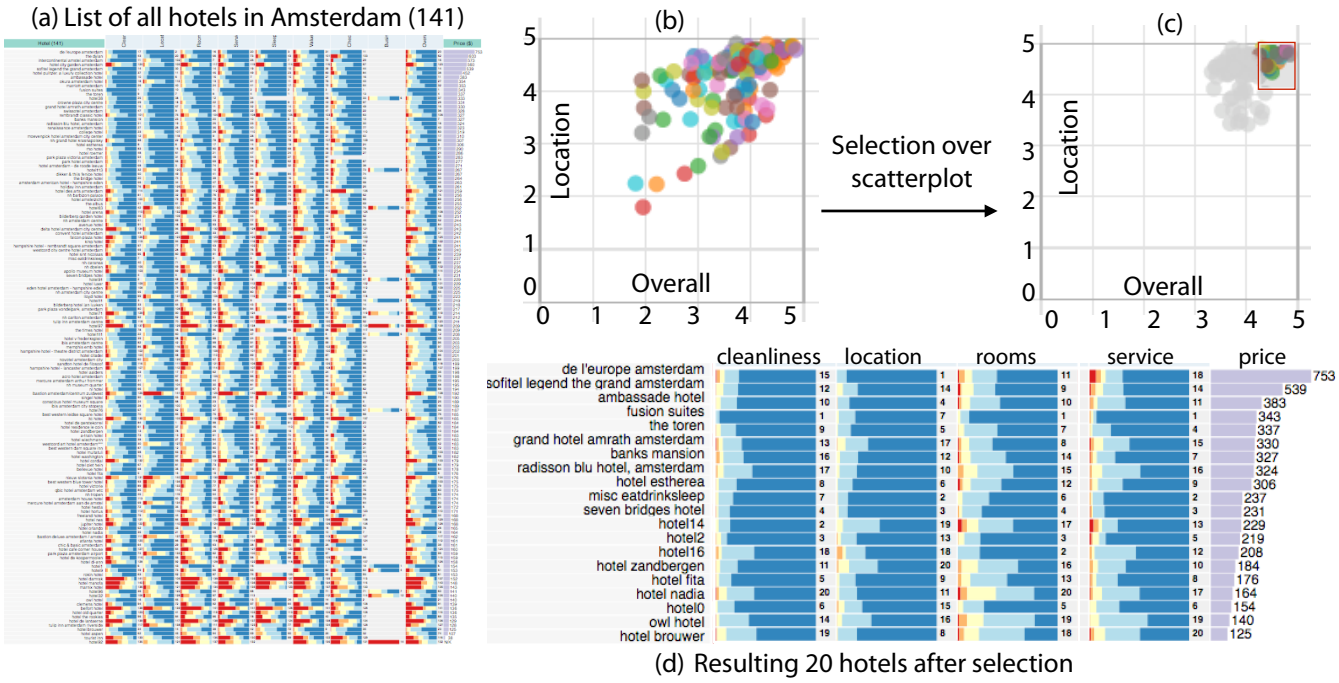


Fig. 6. Selection of hotels in Amsterdam. The list comprises 141 hotels, and can be too long for a user to process (a). One alternative to reduce this list is to focus on specific aspects of interest, such as overall and location. The user selects these aspects and inspect the scatterplot matrix (b). By selecting a rectangular region in the scatterplot (c), the user constrains the list to hotels within the selected region (in this case, the hotels with the highest scores in both aspects). The resulting list has 20 hotels (d), and a new ranking is created using the selected hotels. We observe that the list of hotels has a great price variability while having similar evaluations in the aspects shown, which allows the user to consider several compromises while choosing a hotel.

on how our tool can be useful were shown in the previous section while explaining the interface.

Figure 6 illustrates one possibility of using the system. In this example, we are inspecting for hotel reviews in the city of Amsterdam. The total hotels in this list is 141, which becomes long to establish comparisons among the different aspects and hotels. One way to reduce this list is to apply the selection offered by the interaction with the scatterplot matrix. We configure the creation of the scatterplot matrix in such a way that the user can select the aspects of interest. In this example, we select the location and overall aspects, and inspect the resulting scatterplot. The selection is defined over the scatterplot using a rectangular region, in this case corresponding to the upper-right corner of the scatterplot (hotels with higher scores in the selected aspects). The result of this selection is a list of 20 hotels. It is interesting to observe that some hotels have very different prices but similar stacked bar charts, which means that the user could find a similar service at a lower cost. We believe this process is useful in refining the search to the hotels that satisfy the interests of users.

We include a video in the accompanying material to illustrate the system in action. Other aspects of the interface can be better inspected in the video, such as the many possible orderings using the different aspects, and selection using the scatterplot matrix. We plan to make the prototype publicly available soon in the internet.

## V. Conclusion and Future Work

In this paper, we presented a tool that includes several visualization methods to compare, analyze, and select hotels using the TripAdvisor data as test case. Our goal was to demonstrate how the user interface composed of the visualization using different ranking strategies and selection using the scatterplot matrix of aspects allow comparing hotels.

We plan to continue expanding this work in many different ways. First, we want to conduct an evaluation study with users of different backgrounds to gather feedback on the prototype. It would be very interesting if we could perform this study with an even larger dataset, which would stress test some of the visualizations and selections we implemented. Also, in the current version, we do not show the text of the individual reviews. We want to display reviews when the user selects a specific hotel, but we also consider displaying reviews for multiple hotels. There are many challenges on how to accomplish this, and therefore we deferred this possibility for future work. Another desired feature in our system is to incorporate the time-varying aspect of reviews. This property has a great impact in all the visualizations we considered, since reviews change over time, and therefore all data being displayed is subject to changes throughout time.

## Acknowledgments

We thank the Database and Information Systems Laboratory (DAIS) at the University of Illinois for providing the TripAdvisor reviews database [16], [17].

## REFERENCES


[1] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu, "Opinionseer: Interactive visualization of hotel customer feedback," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1109–1118, November 2010. [Online]. Available: https://www.microsoft.com/en-us/research/publication/opinionseer-interactive-visualization-hotel-customer-feedback/

[2] M. M. Mostafa, "More than words: Social networks text mining for consumer brand sentiments," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4241 – 4251, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413000328

[3] C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang, "Sentiview: Sentiment analysis and visualization for internet popular topics," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 620–630, Nov 2013.

[4] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *2010 IEEE Symposium on Visual Analytics Science and Technology*, Oct 2010, pp. 115–122.

[5] R. Sundberg, A. Eriksson, J. Bini, and P. Nugues, "Visualizing sentiment analysis on a user forum," in *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), 2012, pp. 3573–3579.

[6] B. Alper, H. Yang, E. Haber, and E. Kandogan, "Opinionblocks: Visualizing consumer reviews," *IEEE visWeek 2011 Workshop on Interactive Visual Text Analytics for Decision Making*, Nov 2011.

[7] D. Oelke, M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L. E. Haug, and H. Janetzko, "Visual opinion analysis of customer feedback data," in *2009 IEEE Symposium on Visual Analytics Science and Technology*, Oct 2009, pp. 187–194.

[8] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, and C. Weaver, "Visual analysis of conflicting opinions," in *2006 IEEE Symposium On Visual Analytics Science And Technology*, Oct 2006, pp. 59–66.

[9] V. Setlur and M. C. Stone, "A linguistic approach to categorical color assignment for data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 698–707, Jan 2016.

[10] J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1224–1231, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2007.70515

[11] I. Demir, C. Dick, and R. Westermann, "Multi-charts for comparative 3d ensemble visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2694–2703, Dec 2014.

[12] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu, "Rankexplorer: Visualization of ranking changes in large time series data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2669–2678, Dec 2012.

[13] S. Okubo, T. Iwakura, and K. Misue, "Trend analysis tool with simultaneous visualization of rank and value," in *2013 17th International Conference on Information Visualisation*, July 2013, pp. 517–522.

[14] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "LineUp: Visual analysis of multi-attribute rankings," *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13)*, vol. 19, no. 12, pp. 2277–2286, 2013.

[15] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4065 – 4074, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413000043

[16] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis on review text data: A rating regression approach," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 783–792. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835903

[17] ——, "Latent aspect rating analysis without aspect keyword supervision," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 618–626. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020505