Understanding User Behavior in Online Feedback Reporting

Arjun Talwar
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Lab
Lausanne, Switzerland
arjun@math.stanford.edu

Radu Jurca
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Lab
Lausanne, Switzerland
radu.jurca@epfl.ch

Boi Faltings
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Lab
Lausanne, Switzerland
boi.faltings@epfl.ch

ABSTRACT

Online reviews have become increasingly popular as a way to judge the quality of various products and services. Previous work has demonstrated that contradictory reporting and underlying user biases make judging the true worth of a service difficult. In this paper, we investigate underlying factors that influence user behavior when reporting feedback. We look at two sources of information besides numerical ratings: linguistic evidence from the textual comment accompanying a review, and patterns in the time sequence of reports. We first show that groups of users who amply discuss a certain feature are more likely to agree on a common rating for that feature. Second, we show that a user's rating partly reflects the difference between true quality and prior expectation of quality as inferred from previous reviews. Both give us a less noisy way to produce rating estimates and reveal the reasons behind user bias. Our hypotheses were validated by statistical evidence from hotel reviews on the TripAdvisor website.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Economics, Experimentation, Reliability

Keywords

Online Reviews, Reputation Mechanisms

1. MOTIVATIONS

The spread of the internet has made it possible for online feedback forums (or reputation mechanisms) to become an important channel for *Word-of-mouth* regarding products, services or other types of commercial interactions. Numerous empirical studies [10, 15, 13, 5] show that buyers seriously consider online feedback when making purchasing decisions, and are willing to pay *reputation premiums* for products or services that have a good reputation.

Recent analysis, however, raises important questions regarding the ability of existing forums to reflect the real quality of a product. In the absence of clear incentives, users with a moderate outlook will not bother to voice their opinions, which leads to an unrepresentative sample of reviews. For example, [12, 1] show that Amazon[1] ratings of books or CDs follow with great probability bi-modal, U-shaped distributions where most of the ratings are either very good, or very bad. Controlled experiments, on the other hand, reveal opinions on the same items that are normally distributed. Under these circumstances, using the arithmetic mean to predict quality (as most forums actually do) gives the typical user an estimator with high variance that is often false.

Improving the way we aggregate the information available from online reviews requires a deep understanding of the underlying factors that bias the rating behavior of users. Hu et al. [12] propose the "Brag-and-Moan Model" where users rate only if their utility of the product (drawn from a normal distribution) falls outside a median interval. The authors conclude that the model explains the empirical distribution of reports, and offers insights into smarter ways of estimating the true quality of the product.

In the present paper we extend this line of research, and attempt to explain further facts about the behavior of users when reporting online feedback. Using actual hotel reviews from the TripAdvisor[2] website, we consider two additional sources of information besides the basic numerical ratings submitted by users. The first is simple linguistic evidence from the textual review that usually accompanies the numerical ratings. We use text-mining techniques similar to [7] and [3], however, we are only interested in identifying *what* aspects of the service the user is discussing, without computing the semantic orientation of the text. We find that users who comment more on the same feature are more likely to agree on a common numerical rating for that particular feature. Intuitively, lengthy comments reveal the importance of the feature to the user. Since people tend to be more knowledgeable in the aspects they consider important, users who discuss a given feature in more details might be assumed to have more *authority* in evaluating that feature.

Second we investigate the relationship between a review

[1] http://www.amazon.com
[2] http://www.tripadvisor.com/

EC'07, June 11–15, 2007, San Diego, California, USA.

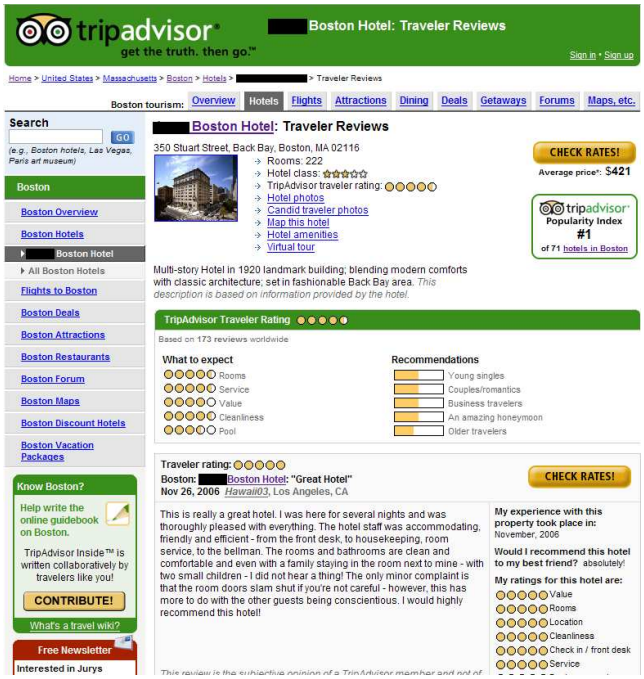

Figure 1: The TripAdvisor page displaying reviews for a popular Boston hotel. Name of hotel and advertisements were deliberatively erased.

and the reviews that preceded it. A perusal of online reviews shows that ratings are often part of discussion threads, where one post is not necessarily independent of other posts. One may see, for example, users who make an effort to contradict, or vehemently agree with, the remarks of previous users. By analyzing the time sequence of reports, we conclude that past reviews influence the future reports, as they create some prior expectation regarding the quality of service. The subjective perception of the user is influenced by the gap between the prior expectation and the actual performance of the service [17, 18, 16, 21] which will later reflect in the user's rating. We propose a model that captures the dependence of ratings on prior expectations, and validate it using the empirical data we collected.

Both results can be used to improve the way reputation mechanisms aggregate the information from individual reviews. Our first result can be used to determine a feature-by-feature estimate of quality, where for each feature, a different subset of reviews (i.e., those with lengthy comments of that feature) is considered. The second leads to an algorithm that outputs a more precise estimate of the real quality.

2. THE DATA SET

We use in this paper real hotel reviews collected from the popular travel site TripAdvisor. TripAdvisor indexes hotels from cities across the world, along with reviews written by travelers. Users can search the site by giving the hotel's name and location (optional). The reviews for a given hotel are displayed as a list (ordered from the most recent to the oldest), with 5 reviews per page. The reviews contain:

- information about the author of the review (e.g., dates of stay, username of the reviewer, location of the reviewer);
- the overall rating (from 1, lowest, to 5, highest);
- a textual review containing a title for the review, free comments, and the main things the reviewer liked and disliked;
- numerical ratings (from 1, lowest, to 5, highest) for different features (e.g., cleanliness, service, location, etc.)

Below the name of the hotel, TripAdvisor displays the address of the hotel, general information (number of rooms, number of stars, short description, etc), the average overall rating, the TripAdvisor ranking, and an average rating for each feature. Figure 1 shows the page for a popular Boston hotel whose name (along with advertisements) was explicitly erased.

We selected three cities for this study: Boston, Sydney and Las Vegas. For each city we considered all hotels that had at least 10 reviews, and recorded all reviews. Table 1 presents the number of hotels considered in each city, the total number of reviews recorded for each city, and the distribution of hotels with respect to the star-rating (as available on the TripAdvisor site). Note that not all hotels have a star-rating.

Table 1: A summary of the data set.

City	# Reviews	# Hotels	# of Hotels with 1,2,3,4 & 5 stars
Boston	3993	58	1+3+17+15+2
Sydney	1371	47	0+0+9+13+10
Las Vegas	5593	40	0+3+10+9+6

For each review we recorded the overall rating, the textual review (title and body of the review) and the numerical rating on 7 features: *Rooms*(R), *Service*(S), *Cleanliness*(C), *Value*(V), *Food*(F), *Location*(L) and *Noise*(N). TripAdvisor does not require users to submit anything other than the overall rating, hence a typical review rates few additional features, regardless of the discussion in the textual comment. Only the features *Rooms*(R), *Service*(S), *Cleanliness*(C) and *Value*(V) are rated by a significant number of users. However, we also selected the features *Food*(F), *Location*(L) and *Noise*(N) because they are referred to in a significant number of textual comments. For each feature we record the numerical rating given by the user, or 0 when the rating is missing. The typical length of the textual comment amounts to approximately 200 words. All data was collected by crawling the TripAdvisor site in September 2006.

2.1 Formal notation

We will formally refer to a review by a tuple (r, T) where:

- $r = (r_f)$ is a vector containing the ratings $r_f \in \{0, 1, \ldots 5\}$ for the features $f \in F = \{O, R, S, C, V, F, L, N\}$; note that the overall rating, r_O, is abusively recorded as the rating for the feature *Overall*(O);
- T is the textual comment that accompanies the review.

Reviews are indexed according to the variable i, such that (r^i, T^i) is the i^{th} review in our database. Since we don't record the username of the reviewer, we will also say that the i^{th} review in our data set was submitted by user i. When we need to consider only the reviews of a given hotel, h, we will use $(r^{i(h)}, T^{i(h)})$ to denote the i^{th} review about the hotel h.

3. EVIDENCE FROM TEXTUAL COMMENTS

The free textual comments associated to online reviews are a valuable source of information for understanding the reasons behind the numerical ratings left by the reviewers. The text may, for example, reveal concrete examples of aspects that the user liked or disliked, thus justifying some of the high, respectively low ratings for certain features. The text may also offer guidelines for understanding the preferences of the reviewer, and the weights of different features when computing an overall rating.

The problem, however, is that free textual comments are difficult to read. Users are required to scroll through many reviews and read mostly repetitive information. Significant improvements would be obtained if the reviews were automatically interpreted and aggregated. Unfortunately, this seems a difficult task for computers since human users often use witty language, abbreviations, cultural specific phrases, and the figurative style.

Nevertheless, several important results use the textual comments of online reviews in an automated way. Using well established natural language techniques, reviews or parts of reviews can be classified as having a positive or negative *semantic orientation*. Pang et al. [2] classify movie reviews into positive/negative by training three different classifiers (Naive Bayes, Maximum Entropy and SVM) using classification features based on unigrams, bigrams or part-of-speech tags.

Dave et al. [4] analyze reviews from CNet and Amazon, and surprisingly show that classification features based on unigrams or bigrams perform better than higher-order *n-grams*. This result is challenged by Cui et al. [3] who look at large collections of reviews crawled from the web. They show that the size of the data set is important, and that bigger training sets allow classifiers to successfully use more complex classification features based on *n-grams*.

Hu and Liu [11] also crawl the web for product reviews and automatically identify product attributes that have been discussed by reviewers. They use Wordnet to compute the semantic orientation of product evaluations and summarize user reviews by extracting positive and negative evaluations of different product features. Popescu and Etzioni [20] analyze a similar setting, but use search engine hit-counts to identify product attributes; the semantic orientation is assigned through the *relaxation labeling technique*.

Ghose et al. [7, 8] analyze seller reviews from the Amazon secondary market to identify the different dimensions (e.g., delivery, packaging, customer support, etc.) of reputation. They parse the text, and tag the part-of-speech for each word. Frequent nouns, noun phrases and verbal phrases are identified as dimensions of reputation, while the corresponding *modifiers* (i.e., adjectives and adverbs) are used to derive numerical scores for each dimension. The enhanced reputation measure correlates better with the pricing information observed in the market. Pavlou and Dimoka [19] analyze eBay reviews and find that textual comments have an important impact on reputation premiums.

Our approach is similar to the previously mentioned works, in the sense that we identify the aspects (i.e., hotel features) discussed by the users in the textual reviews. However, we do not compute the semantic orientation of the text, nor attempt to infer missing ratings.

We define the *weight*, w_f^i, of feature $f \in F$ in the text T^i associated with the review (r^i, T^i), as the fraction of T^i dedicated to discussing aspects (both positive and negative) related to feature f. We propose an elementary method to approximate the values of these weights. For each feature we manually construct the word list L_f containing approximately 50 words that are most commonly associated to the feature f. The initial words were selected from reading some of the reviews, and seeing what words coincide with discussion of which features. The list was then extended by adding all thesaurus entries that were related to the initial words. Finally, we brainstormed for missing words that would normally be associated with each of the features.

Let $L_f \cap T^i$ be the list of terms common to both L_f and T_i. Each term of L_f is counted the number of times it appears in T^i, with two exceptions:

- in cases where the user submits a title to the review, we account for the title text by appending it three times to the review text T^i. The intuitive assumption is that the user's opinion is more strongly reflected in the title, rather than in the body of the review. For example, many reviews are accurately summarized by titles such as *"Excellent service, terrible location"* or *"Bad value for money"*;

- certain words that occur only once in the text are counted multiple times if their relevance to that feature is particularly strong. These were 'root' words for each feature (e.g., 'staff' is a root word for the feature *Service*), and were weighted either 2 or 3. Each feature was assigned up to 3 such root words, so almost all words are counted only once.

The list of words for the feature Rooms is given for reference in Appendix A.

The weight w_f^i is computed as:

$$w_f^i = \frac{|L_f \cap T^i|}{\sum_{f \in F} |L_f \cap T^i|} \quad (1)$$

where $|L_f \cap T^i|$ is the number of terms common to L_f and T^i. The weight for the feature *Overall* was set to $\min\{\frac{|T^i|}{5000}, 1\}$ where $|T^i|$ is the number of character in T^i.

The following is a TripAdvisor review for a Boston hotel (the name of the hotel is omitted): *"I'll start by saying that I'm more of a Holiday Inn person than a *** type. So I get frustrated when I pay double the room rate and get half the amenities that I'd get at a Hampton Inn or Holiday Inn. The location was definitely the main asset of this place. It was only a few blocks from the Hynes Center subway stop and it was easy to walk to some good restaurants in the Back Bay area. Boylston isn't far off at all. So I had no trouble with foregoing a rental car and taking the subway from the airport to the hotel and using the subway for any other travel. Otherwise, they make you pay for anything and everything.*

And when you've already dropped $215/night on the room, that gets frustrating.The room itself was decent, about what I would expect. Staff was also average, not bad and not excellent. Again, I think you're paying for location and the ability to walk to a lot of good stuff. But I think next time I'll stay in Brookline, get more amenities, and use the subway a bit more.

This numerical ratings associated to this review are $r_O = 3$, $r_R = 3$, $r_S = 3$, $r_C = 4$, $r_V = 2$ for features *Overall*(O), *Rooms*(R), *Service*(S), *Cleanliness*(C) and *Value*(V) respectively. The ratings for the features *Food*(F), *Location*(L) and *Noise*(N) are absent (i.e., $r_F = r_L = r_N = 0$).

The weights w_f are computed from the following lists of common terms:
$L_R \cap T$ ={room}; $w_R = 0.066$
$L_S \cap T$ ={3 * Staff, amenities}; $w_S = 0.267$
$L_C \cap T = \emptyset$; $w_C = 0$
$L_V \cap T$ ={$, rate}; $w_V = 0.133$
$L_F \cap T$ ={restaurant}; $w_F = 0.067$
$L_L \cap T$ ={2 * center, 2 * walk, 2 * location, area}; $w_L = 0.467$
$L_N \cap T = \emptyset$; $w_N = 0$
The root words 'Staff' and 'Center' were tripled and doubled respectively. The overall weight of the textual review is $w_O = 0.197$. These values account reasonably well for the weights of different features in the discussion of the reviewer.

One point to note is that some terms in the lists L_f possess an inherent semantic orientation. For example the word 'grime' (belonging to the list L_C) would be used most often to assert the presence, and not the absence of grime. This is unavoidable, but care was taken to ensure words from both sides of the spectrum were used. For this reason, some lists such as L_R contain only nouns of objects that one would typically describe in a room (see Appendix A).

The goal of this section is to analyse the influence of the weights w_f^i on the numerical ratings r_f^i. Intuitively, users who spent a lot of their time discussing a feature f (i.e., w_f^i is high) had something to say about their experience with regard to this feature. Obviously, feature f is important for user i. Since people tend to be more knowledgeable in the aspects they consider important, our hypothesis is that the ratings r_f^i (corresponding to high weights w_f^i) constitute a subset of "expert" ratings for feature f.

Figure 2 plots the distribution of the rates $r_C^{i(h)}$ with respect to the weights $w_C^{i(h)}$ for the cleanliness of a Las Vegas hotel, h. Here, the high ratings are restricted to the reviews that discuss little the cleanliness. Whenever cleanliness appears in the discussion, the ratings are low. Many hotels exhibit similar rating patterns for various features. Ratings corresponding to low weights span the whole spectrum from 1 to 5, while the ratings corresponding to high weights are more grouped together (either around good or bad ratings).

We therefore make the following hypothesis:

HYPOTHESIS 1. *The ratings r_f^i corresponding to the reviews where w_f^i is high, are more similar to each other than to the overall collection of ratings.*

To test the hypothesis, we take the entire set of reviews, and feature by feature, we compute the standard deviation of the ratings with high weights, and the standard deviation of the entire set of ratings. High weights were defined as those belonging to the upper 20% of the weight range for the corresponding feature. If Hypothesis 1 were true, the standard deviation of all ratings should be higher than the standard deviation of the ratings with high weights.

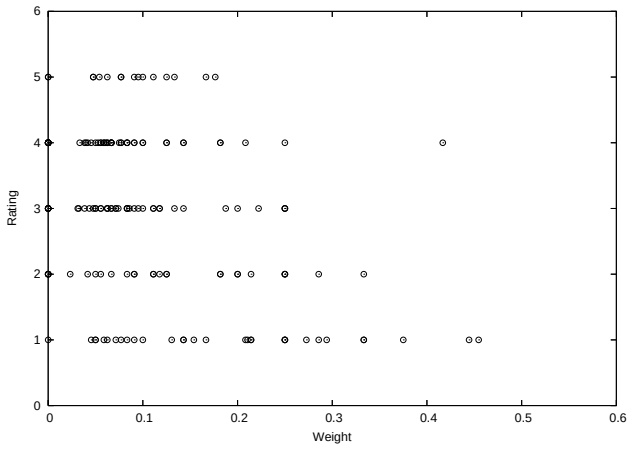

Figure 2: The distribution of ratings against the weight of the cleanliness feature.

We use a standard T-test to measure the significance of the results. City by city and feature by feature, Table 2 presents the average standard deviation of all ratings, and the average standard deviation of ratings with high weights. Indeed, the ratings with high weights have lower standard deviation, and the results are significant at the standard 0.05 significance threshold (although for certain cities taken independently there doesn't seem to be a significant difference, the results are significant for the entire data set). Please note that only the features O,R,S,C and V were considered, since for the others (F, L, and N) we didn't have enough ratings.

Table 2: Average standard deviation for *all* ratings, and average standard deviation for ratings with *high* weights. In square brackets, the corresponding p-values for a positive difference between the two.

City	O	R	S	C	V
all	1.189	0.998	1.144	0.935	1.123
Boston high	0.948	0.778	0.954	0.767	0.891
p-val	[0.000]	[0.004]	[0.045]	[0.080]	[0.009]
all	1.040	0.832	1.101	0.847	0.963
Sydney high	0.801	0.618	0.691	0.690	0.798
p-val	[0.012]	[0.023]	[0.000]	[0.377]	[0.037]
all	1.272	1.142	1.184	1.119	1.242
Vegas high	1.072	0.752	1.169	0.907	1.003
p-val	[0.0185]	[0.001]	[0.918]	[0.120]	[0.126]

Hypothesis 1 not only provides some basic understanding regarding the rating behavior of online users, it also suggests some ways of computing better quality estimates. We can, for example, construct a feature-by-feature quality estimate with much lower variance: for each feature we take the subset of reviews that amply discuss that feature, and output as a quality estimate the average rating for this subset. Initial experiments suggest that the average feature-by-feature ratings computed in this way are different from the average ratings computed on the whole data set. Given that, indeed, high weights are indicators of "expert" opinions, the estimates obtained in this way are more accurate than the current ones. Nevertheless, the validation of this underlying assumption requires further controlled experiments.

4. THE INFLUENCE OF PAST RATINGS

Two important assumptions are generally made about reviews submitted to online forums. The first is that ratings truthfully reflect the quality observed by the users; the second is that reviews are independent from one another. While anecdotal evidence [9, 22] challenges the first assumption[3], in this section, we address the second.

A perusal of online reviews shows that reviews are often part of discussion threads, where users make an effort to contradict, or vehemently agree with the remarks of previous users. Consider, for example, the following review:
*"I don't understand the negative reviews... the hotel was a little dark, but that was the style. It was very artsy. Yes it was close to the freeway, but in my opinion the sound of an occasional loud car is better than hearing the "ding ding" of slot machines all night! The staff on-hand is FABULOUS. The waitresses are great (and *** does not deserve the bad review she got, she was 100% attentive to us!), the bartenders are friendly and professional at the same time..."*

Here, the user was disturbed by previous negative reports, addressed these concerns, and set about trying to correct them. Not surprisingly, his ratings were considerably higher than the average ratings up to this point.

It seems that TripAdvisor users regularly read the reports submitted by previous users before booking a hotel, or before writing a review. Past reviews create some prior expectation regarding the quality of service, and this expectation has an influence on the submitted review. We believe this observation holds for most online forums. The subjective perception of quality is directly proportional to how well the actual experience meets the prior expectation, a fact confirmed by an important line of econometric and marketing research [17, 18, 16, 21].

The correlation between the reviews has also been confirmed by recent research on the dynamics of online review forums [6].

4.1 Prior Expectations

We define the prior expectation of user i regarding the feature f, as the average of the previously available ratings on the feature f[4]:

$$e_f(i) = \frac{\sum_{j<i, r_f^j \neq 0} r_f^j}{\sum_{j<i, r_f^j \neq 0} 1}$$

As a first hypothesis, we assert that the rating r_f^i is a function of the prior expectation $e_f(i)$:

HYPOTHESIS 2. *For a given hotel and feature, given the reviews i and j such that $e_f(i)$ is high and $e_f(j)$ is low, the rating r_f^j exceeds the rating r_f^i.*

We define *high* and *low* expectations as those that are above, respectively below a certain cutoff value θ. The set of reviews preceded by high, respectively low expectations

[3]part of Amazon reviews were recognized as strategic posts by book authors or competitors

[4]if no previous ratings were assigned for feature f, $e_f(i)$ is assigned a default value of 4.

Table 3: Average ratings for reviews preceded by low (first value in the cell) and high (second value in the cell) expectations. The P-values for a positive difference are given square brackets.

City	O	R	S	C	V
Boston	3.953 3.364 [0.011]	4.045 3.590 [0.028]	3.985 3.485 [0.0086]	4.252 3.641 [0.0168]	3.946 3.242 [0.0034]
Sydney	4.284 3.756 [0.000]	4.358 3.537 [0.000]	4.064 3.436 [0.035]	4.530 3.918 [0.009]	4.428 3.495 [0.000]
Las Vegas	3.494 3.140 [0.190]	3.674 3.530 [0.529]	3.713 2.952 [0.007]	3.689 3.530 [0.529]	3.580 3.351 [0.253]

are defined as follows:

$$R_f^{high} = \{r_f^i | e_f(i) > \theta\}$$
$$R_f^{low} = \{r_f^i | e_f(i) < \theta\}$$

These sets are specific for each (hotel, feature) pair, and in our experiments we took $\theta = 4$. This rather high value is close to the average rating across all features across all hotels, and is justified by the fact that our data set contains mostly high quality hotels.

For each city, we take all hotels and compute the average ratings in the sets R_f^{high} and R_f^{low} (see Table 3). The average rating amongst reviews following low prior expectations is significantly higher than the average rating following high expectations.

As further evidence, we consider all hotels for which the function $e_V(i)$ (the expectation for the feature *Value*) has a high value (greater than 4) for some i, and a low value (less than 4) for some other i. Intuitively, these are the hotels for which there is a minimal degree of variation in the timely sequence of reviews: i.e., the cumulative average of ratings was at some point high and afterwards became low, or vice-versa. Such variations are observed for about half of all hotels in each city. Figure 3 plots the median (across considered hotels) rating, r_V, when $e_f(i)$ is not more than x but greater than $x - 0.5$.

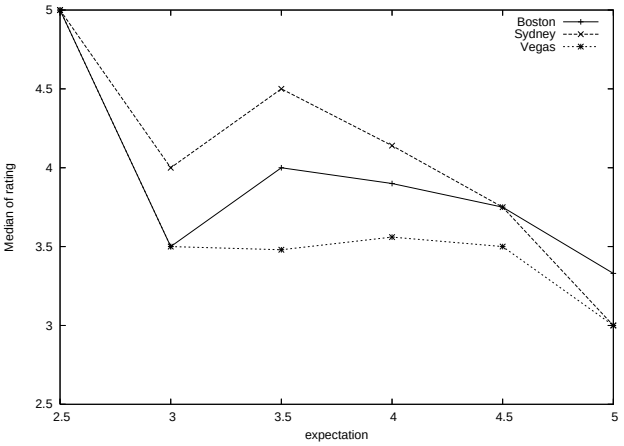

Figure 3: The ratings tend to decrease as the expectation increases.

There are two ways to interpret the function $e_f(i)$:

- The expected value for feature f obtained by user i before his experience with the service, acquired by reading reports submitted by past users. In this case, an overly high value for $e_f(i)$ would drive the user to submit a negative report (or vice versa), stemming from the difference between the actual value of the service, and the inflated expectation of this value acquired before his experience.
- The expected value of feature f for all subsequent visitors of the site, if user i were not to submit a report. In this case, the motivation for a negative report following an overly high value of e_f is different: user i seeks to *correct* the expectation of future visitors to the site. Unlike the interpretation above, this does not require the user to derive an *a priori* expectation for the value of f.

Note that neither interpretation implies that the average up to report i is inversely related to the rating at report i. There might exist a measure of influence exerted by past reports that pushes the user behind report i to submit ratings which to some extent conforms with past reports: a low value for $e_f(i)$ can influence user i to submit a low rating for feature f because, for example, he fears that submitting a high rating will make him out to be a person with low standards[5]. This, at first, appears to contradict Hypothesis 2. However, this conformity rating cannot continue indefinitely: once the set of reports project a sufficiently deflated estimate for v_f, future reviewers with comparatively positive impressions will seek to correct this misconception.

4.2 Impact of textual comments on quality expectation

Further insight into the rating behavior of TripAdvisor users can be obtained by analyzing the relationship between the weights w_f and the values $e_f(i)$. In particular, we examine the following hypothesis:

HYPOTHESIS 3. *When a large proportion of the text of a review discusses a certain feature, the difference between the rating for that feature and the average rating up to that point tends to be large.*

The intuition behind this claim is that when the user is adamant about voicing his opinion regarding a certain feature, his opinion differs from the collective opinion of previous postings. This relies on the characteristic of reputation systems as feedback forums where a user is interested in projecting his opinion, with particular strength if this opinion differs from what he perceives to be the general opinion.

To test Hypothesis 3 we measure the average absolute difference between the expectation $e_f(i)$ and the rating r_f^i when the weight w_f^i is high, respectively low. Weights are classified high or low by comparing them with certain cutoff values: w_f^i is low if smaller than 0.1, while w_f^i is high if greater than θ_f. Different cutoff values were used for different features: $\theta_R = 0.4$, $\theta_S = 0.4$, $\theta_C = 0.2$, and $\theta_V = 0.7$. *Cleanliness* has a lower cutoff since it is a feature rarely discussed; *Value* has a high cutoff for the opposite reason. Results are presented in Table 4.

[5]The idea that negative reports can encourage further negative reporting has been suggested before [14]

Table 4: Average of $|r_f^i - e_f(i)|$ when weights are high (first value in the cell) and low (second value in the cell) with P-values for the difference in sq. brackets.

City	R	S	C	V
Boston	1.058 0.701 [0.022]	1.208 0.838 [0.063]	1.728 0.760 [0.000]	1.356 0.917 [0.218]
Sydney	1.048 0.752 [0.179]	1.351 0.759 [0.009]	1.218 0.767 [0.165]	1.318 0.908 [0.495]
Las Vegas	1.184 0.772 [0.071]	1.378 0.834 [0.020]	1.472 0.808 [0.006]	1.642 1.043 [0.076]

This demonstrates that when weights are unusually high, users tend to express an opinion that does not conform to the net average of previous ratings. As we might expect, for a feature that rarely was a high weight in the discussion, (e.g., cleanliness) the difference is particularly large. Even though the difference in the feature *Value* is quite large for Sydney, the P-value is high. This is because only few reviews discussed value heavily. The reason could be cultural or because there was less of a reason to discuss this feature.

4.3 Reporting Incentives

Previous models suggest that users who are not highly opinionated will not choose to voice their opinions [12]. In this section, we extend this model to account for the influence of expectations. The motivation for submitting feedback is not only due to extreme opinions, but also to the difference between the current reputation (i.e., the prior expectation of the user) and the actual experience.

Such a rating model produces ratings that most of the time deviate from the current average rating. The ratings that confirm the prior expectation will rarely be submitted. We test on our data set the proportion of ratings that attempt to "correct" the current estimate. We define a deviant rating as one that deviates from the current expectation by at least some threshold θ, i.e., $|r_f^i - e_f(i)| \geq \theta$. For each of the three considered cities, the following tables, show the proportion of deviant ratings for $\theta = 0.5$ and $\theta = 1$.

Table 5: Proportion of deviant ratings with $\theta = 0.5$

City	O	R	S	C	V
Boston	0.696	0.619	0.676	0.604	0.684
Sydney	0.645	0.615	0.672	0.614	0.675
Las Vegas	0.721	0.641	0.694	0.662	0.724

Table 6: Proportion of deviant ratings with $\theta = 1$

City	O	R	S	C	V
Boston	0.420	0.397	0.429	0.317	0.446
Sydney	0.360	0.367	0.442	0.336	0.489
Las Vegas	0.510	0.421	0.483	0.390	0.472

The above results suggest that a large proportion of users (close to one half, even for the high threshold value $\theta = 1$) deviate from the prior average. This reinforces the idea that users are more likely to submit a report when they believe they have something distinctive to add to the current stream of opinions for some feature. Such conclusions are in total agreement with prior evidence that the distribution of reports often follows bi-modal, U-shaped distributions.

5. MODELLING THE BEHAVIOR OF RATERS

To account for the observations described in the previous sections, we propose a model for the behavior of the users when submitting online reviews. For a given hotel, we make the assumption that the quality experienced by the users is normally distributed around some value v_f, which represents the "objective" quality offered by the hotel on the feature f. The rating submitted by user i on feature f is:

$$\hat{r}_f^i = \delta_f v_f^i + (1-\delta_f) \cdot sign\Big(v_f^i - e_f(i)\Big)\Big[c + d(v_f^i, e_f(i)|w_f^i)\Big] \quad (2)$$

where:

- v_f^i is the (unknown) quality actually experienced by the user. v_f^i is assumed normally distributed around some value v_f;
- $\delta_f \in [0,1]$ can be seen as a measure of the bias when reporting feedback. High values reflect the fact that users rate objectively, without being influenced by prior expectations. The value of δ_f may depend on various factors; we fix one value for each feature f;
- c is a constant between 1 and 5;
- w_f^i is the weight of feature f in the textual comment of review i, computed according to Eq. (1);
- $d(v_f^i, e_f(i)|w_f^i)$ is a distance function between the expectation and the observation of user i. The distance function satisfies the following properties:
  - $d(y,z|w) \geq 0$ for all $y,z \in [0,5]$, $w \in [0,1]$;
  - $|d(y,z|w)| < |d(z,x|w)|$ if $|y-z| < |z-x|$;
  - $|d(y,z|w_1)| < |d(y,z|w_2)|$ if $w_1 < w_2$;
  - $c + d(v_f, e_f(i)|w_f^i) \in [1,5]$;

  The second term of Eq. (2) encodes the bias of the rating. The higher the distance between the true observation v_f^i and the function e_f, the higher the bias.

5.1 Model Validation

We use the data set of TripAdvisor reviews to validate the behavior model presented above. We split for convenience the rating values in three ranges: bad ($B = \{1,2\}$), indifferent ($I = \{3,4\}$), and good ($G = \{5\}$), and perform the following two tests:

- First, we will use our model to predict the ratings that have extremal values. For every hotel, we take the sequence of reports, and whenever we encounter a rating that is either good or bad (but not indifferent) we try to predict it using Eq. (2)
- Second, instead of predicting the value of extremal ratings, we try to classify them as either good or bad. For every hotel we take the sequence of reports, and for each report (regardless of it value) we classify it as being good or bad

However, to perform these tests, we need to estimate the objective value, v_f, that is the average of the true quality observations, v_f^i. The algorithm we are using is based on the intuition that the amount of conformity rating is minimized. In other words, the value v_f should be such that as often as possible, bad ratings follow expectations above v_f and good ratings follow expectations below v_f.

Formally, we define the sets:

$$\Gamma_1 = \{i | e_f(i) < v_f \text{ and } r_f^i \in B\};$$
$$\Gamma_2 = \{i | e_f(i) > v_f \text{ and } r_f^i \in G\};$$

that correspond to irregularities where even though the expectation at point i is lower than the delivered value, the rating is poor, and vice versa. We define v_f as the value that minimize these union of the two sets:

$$v_f = \arg\min_{v_f} |\Gamma_1 \cup \Gamma_2| \quad (3)$$

In Eq. (2) we replace v_f^i by the value v_f computed in Eq. (3), and use the following distance function:

$$d(v_f, e_f(i)|w_f^i) = \frac{|v_f - e_f(i)|}{v_f - e_f(i)} \sqrt{|{v_f}^2 - e_f(i)^2|} \cdot (1 + 2w_f^i);$$

The constant $c \in I$ was set to $\min\{\max\{e_f(i), 3\}\}, 4\}$. The values for δ_f were fixed at $\{0.7, 0.7, 0.8, 0.7, 0.6\}$ for the features {*Overall*, *Rooms*, *Service*, *Cleanliness*, *Value*} respectively. The weights are computed as described in Section 3.

As a first experiment, we take the sets of "extremal" ratings $\{r_f^i | r_f^i \notin I\}$ for each hotel and feature. For every such rating, r_f^i, we try to estimate it by computing $\hat{r}_f^i$ using Eq. (2). We compare this estimator with the one obtained by simply averaging the ratings over all hotels and features: i.e.,

$$\bar{r}_f = \frac{\sum_{j, r_f^j \neq 0} r_f^j}{\sum_{j, r_f^j \neq 0} 1};$$

Table 7 presents the ratio between the root mean square error (RMSE) when using $\hat{r}_f^i$ and $\bar{r}_f$ to estimate the actual ratings. In all cases the estimate produced by our model is better than the simple average.

Table 7: Average of $\frac{RMSE(\hat{r}_f)}{RMSE(\bar{r}_f)}$

City	O	R	S	C	V
Boston	0.987	0.849	0.879	0.776	0.913
Sydney	0.927	0.817	0.826	0.720	0.681
Las Vegas	0.952	0.870	0.881	0.947	0.904

As a second experiment, we try to distinguish the sets $B_f = \{i | r_f^i \in B\}$ and $G_f = \{i | r_f^i \in G\}$ of bad, respectively good ratings on the feature f. For example, we compute the set B_f using the following classifier (called σ):

$$r_f^i \in B_f \ (\sigma_f(i) = 1) \Leftrightarrow \hat{r}_f^i \leq 4;$$

Tables 8, 9 and 10 present the Precision(p), Recall(r) and $s = \frac{2pr}{p+r}$ for classifier σ, and compares it with a naive majority classifier, τ, $\tau_f(i) = 1 \Leftrightarrow |B_f| \geq |G_f|$:

We see that recall is always higher for σ and precision is usually slightly worse. For the s metric σ tends to add a

Table 8: Precision(p), Recall(r), s=$\frac{2pr}{p+r}$ while spotting poor ratings for Boston

		O	R	S	C	V
σ	p	0.678	0.670	0.573	0.545	0.610
	r	0.626	0.659	0.619	0.612	0.694
	s	0.651	0.665	0.595	0.577	0.609
τ	p	0.684	0.706	0.647	0.611	0.633
	r	0.597	0.541	0.410	0.383	0.562
	s	0.638	0.613	0.502	0.471	0.595

Table 9: Precision(p), Recall(r), s=$\frac{2pr}{p+r}$ while spotting poor ratings for Las Vegas

		O	R	S	C	V
σ	p	0.654	0.748	0.592	0.712	0.583
	r	0.608	0.536	0.791	0.474	0.610
	s	0.630	0.624	0.677	0.569	0.596
τ	p	0.685	0.761	0.621	0.748	0.606
	r	0.542	0.505	0.767	0.445	0.441
	s	0.605	0.607	0.670	0.558	0.511

1-20% improvement over τ, much higher in some cases for hotels in Sydney. This is likely because Sydney reviews are more positive than those of the American cities and cases where the number of bad reviews exceeded the number of good ones are rare. Replacing the test algorithm with one that plays a 1 with probability equal to the proportion of bad reviews improves its results for this city, but it is still outperformed by around 80%.

6. SUMMARY OF RESULTS AND CONCLUSION

The goal of this paper is to explore the factors that drive a user to submit a particular rating, rather than the incentives that encouraged him to submit a report in the first place. For that we use two additional sources of information besides the vector of numerical ratings: first we look at the textual comments that accompany the reviews, and second we consider the reports that have been previously submitted by other users.

Using simple natural language processing algorithms, we were able to establish a correlation between the weight of a certain feature in the textual comment accompanying the review, and the noise present in the numerical rating. Specifically, it seems that users who discuss amply a certain feature are likely to agree on a common rating. This observation allows the construction of feature-by-feature estimators of quality that have a lower variance, and are hopefully less noisy. Nevertheless, further evidence is required to support the intuition that ratings corresponding to high weights are *expert* opinions that deserve to be given higher priority when computing estimates of quality.

Second, we emphasize the dependence of ratings on previous reports. Previous reports create an expectation of quality which affects the subjective perception of the user. We validate two facts about the hotel reviews we collected from TripAdvisor: First, the ratings following low expectations (where the expectation is computed as the average of the previous reports) are likely to be higher than the ratings following high expectations. Intuitively, the perception of quality (and consequently the rating) depends on how well the actual experience of the user meets her expectation. Second, we include evidence from the textual comments, and find that when users devote a large fraction of the text to discussing a certain feature, they are likely to motivate a divergent rating (i.e., a rating that does not conform to the prior expectation). Intuitively, this supports the hypothesis that review forums act as discussion groups where users are keen on presenting and motivating their own opinion.

Table 10: Precision(p), Recall(r), s=$\frac{2pr}{p+r}$ while spotting poor ratings for Sydney

		O	R	S	C	V
σ	p	0.650	0.463	0.544	0.550	0.580
	r	0.234	0.378	0.571	0.169	0.592
	s	0.343	0.452	0.557	0.259	0.586
τ	p	0.562	0.615	0.600	0.500	0.600
	r	0.054	0.098	0.101	0.015	0.175
	s	0.098	0.168	0.172	0.030	0.271

We have captured the empirical evidence in a behavior model that predicts the ratings submitted by the users. The final rating depends, as expected, on the true observation, and on the gap between the observation and the expectation. The gap tends to have a bigger influence when an important fraction of the textual comment is dedicated to discussing a certain feature. The proposed model was validated on the empirical data and provides better estimates of the ratings actually submitted.

One assumption that we make is about the existence of an objective quality value v_f for the feature f. This is rarely true, especially over large spans of time. Other explanations might account for the correlation of ratings with past reports. For example, if $e_f(i)$ reflects the true value of f at a point in time, the difference in the ratings following high and low expectations can be explained by hotel revenue models that are maximized when the value is modified accordingly. However, the idea that variation in ratings is not primarily a function of variation in value turns out to be a useful one. Our approach to approximate this elusive 'objective value' is by no means perfect, but conforms neatly to the idea behind the model.

A natural direction for future work is to examine concrete applications of our results. Significant improvements of quality estimates are likely to be obtained by incorporating all empirical evidence about rating behavior. Exactly how different factors affect the decisions of the users is not clear. The answer might depend on the particular application, context and culture.

7. REFERENCES

[1] A. Admati and P. Pfleiderer. Noisytalk.com: Broadcasting opinions in a noisy environment. Working Paper 1670R, Stanford University, 2000.

[2] P. B., L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the EMNLP-02, the Conference on Empirical Methods in Natural Language Processing*, 2002.

[3] H. Cui, V. Mittal, and M. Datar. Comparative

Experiments on Sentiment Classification for Online Product Reviews. In *Proceedings of AAAI*, 2006.

[4] K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery:opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on the World Wide Web (WWW03)*, 2003.

[5] C. Dellarocas, N. Awad, and X. Zhang. Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures. Working paper, 2006.

[6] C. Forman, A. Ghose, and B. Wiesenfeld. A Multi-Level Examination of the Impact of Social Identities on Economic Transactions in Electronic Markets. Available at SSRN: http://ssrn.com/abstract=918978, July 2006.

[7] A. Ghose, P. Ipeirotis, and A. Sundararajan. Reputation Premiums in Electronic Peer-to-Peer Markets: Analyzing Textual Feedback and Network Structure. In *Third Workshop on Economics of Peer-to-Peer Systems, (P2PECON)*, 2005.

[8] A. Ghose, P. Ipeirotis, and A. Sundararajan. The Dimensions of Reputation in electronic Markets. Working Paper CeDER-06-02, New York University, 2006.

[9] A. Harmon. Amazon Glitch Unmasks War of Reviewers. *The New York Times*, February 14, 2004.

[10] D. Houser and J. Wooders. Reputation in Auctions: Theory and Evidence from eBay. *Journal of Economics and Management Strategy*, 15:353–369, 2006.

[11] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD04)*, 2004.

[12] N. Hu, P. Pavlou, and J. Zhang. Can Online Reviews Reveal a Product's True Quality? In *Proceedings of ACM Conference on Electronic Commerce (EC 06)*, 2006.

[13] K. Kalyanam and S. McIntyre. Return on reputation in online auction market. Working Paper 02/03-10-WP, Leavey School of Business, Santa Clara University., 2001.

[14] L. Khopkar and P. Resnick. Self-Selection, Slipping, Salvaging, Slacking, and Stoning: the Impacts of Negative Feedback at eBay. In *Proceedings of ACM Conference on Electronic Commerce (EC 05)*, 2005.

[15] M. Melnik and J. Alm. Does a seller's reputation matter? evidence from ebay auctions. *Journal of Industrial Economics*, 50(3):337–350, 2002.

[16] R. Olshavsky and J. Miller. Consumer Expectations, Product Performance and Perceived Product Quality. *Journal of Marketing Research*, 9:19–21, February 1972.

[17] A. Parasuraman, V. Zeithaml, and L. Berry. A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, 49:41–50, 1985.

[18] A. Parasuraman, V. Zeithaml, and L. Berry. SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, 64:12–40, 1988.

[19] P. Pavlou and A. Dimoka. The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation. *Information Systems Research*, 17(4):392–414, 2006.

[20] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[21] R. Teas. Expectations, Performance Evaluation, and Consumers' Perceptions of Quality. *Journal of Marketing*, 57:18–34, 1993.

[22] E. White. Chatting a Singer Up the Pop Charts. *The Wall Street Journal*, October 15, 1999.

APPENDIX

A. LIST OF WORDS, L_R, ASSOCIATED TO THE FEATURE ROOMS

All words serve as prefixes: *room, space, interior, decor, ambiance, atmosphere, comfort, bath, toilet, bed, building, wall, window, private, temperature, sheet, linen, pillow, hot, water, cold, water, shower, lobby, furniture, carpet, air, condition, mattress, layout, design, mirror, ceiling, lighting, lamp, sofa, chair, dresser, wardrobe, closet*