Beyond the Stars: Improving Rating Predictions using Review Text Content

Gayatree Ganu
Computer Science
Rutgers University
gganu@cs.rutgers.edu

Noémie Elhadad
Biomedical Informatics
Columbia University
noemie@dbmi.columbia.edu

Amélie Marian
Computer Science
Rutgers University
amelie@cs.rutgers.edu

ABSTRACT

Online reviews are an important asset for users deciding to buy a product, see a movie, or go to a restaurant, as well as for businesses tracking user feedback. However, most reviews are written in a free-text format, and are therefore difficult for computer systems to understand, analyze, and aggregate. One consequence of this lack of structure is that searching text reviews is often frustrating for users. User experience would be greatly improved if the structure and sentiment conveyed in the content of the reviews were taken into account. Our work focuses on identifying this information from free-form text reviews, and using the knowledge to improve user experience in accessing reviews. Specifically, we focused on improving recommendation accuracy in a restaurant review scenario. In this paper, we report on our classification effort, and on the insight on user-reviewing behavior that we gained in the process. We propose new ad-hoc and regression-based recommendation measures, that both take into account the textual component of user reviews. Our results show that using textual information results in better general or personalized review score predictions than those derived from the numerical star ratings given by the users.

1. INTRODUCTION

The recent Web 2.0 user-generated content revolution has enabled people to broadcast their knowledge and experience. Web users have whole-heartedly incorporated peer-authored posts into their lives, whether to make purchasing decisions based on recommendations or to plan a night out using restaurant and movie reviews. Despite the growing popularity, there has been little research on the quality of the content. In addition, web sites providing user reviews are surprisingly technologically poor: users often have no choice but to browse through massive amounts of text to find a particular piece of interesting information.

Accessing and searching text reviews is frustrating when users only have a vague idea of the product or its features and they need a recommendation or closest match. Keyword searches typically do not provide good results, as the same keywords routinely appear in good and in bad reviews [1]. Yet another challenge in understanding reviews is that a reviewer's overall rating might be largely reflective of product features in which the search user is not interested. Consider the following example:

EXAMPLE 1: *The New York restaurant Lucky Cheng's in* Citysearch *(http://newyork.citysearch.com) has 65 user reviews of which 40 reviews have a 4 or 5 star rating (out of 5 possible stars). Majority positive reviews, however, praise the ambience of the restaurant, as shown in the following sentences extracted from the reviews:*

- "obviously it's not the food or drinks that is the attraction, but the burlesque show"
- "The food was okay, not great, not bad.[...]Our favorite part, though, was the show!"

The negative reviews complain at length about the price and service. A user not interested in ambience would probably not want to dine at this restaurant. However, a recommendation based on star ratings would label this restaurant as a high-quality restaurant.

User experience would be greatly improved if the structure of the content in reviews was taken into account, i.e., if review parts pertaining to different product features (e.g., food, ambience, price, service for a restaurant), as well as the sentiment of the reviewer towards each feature (e.g., positive, negative or neutral) were identified. This information, coupled with the metadata associated with a product (e.g., location or cuisine for restaurants), can then be used to analyze and access reviews. However, identifying structured information from free-form text is a challenging task as users routinely enter informal text with poor spelling and grammar. We performed an in-depth classification of a real-world restaurant review data set and report on our techniques and findings in this paper.

A typical application over reviews is recommendation systems, in which users do not search reviews directly but are suggested products that would best match some definition of preference [7]. Current recommendation systems, such as the ones used by Netflix or Amazon, rely predominantly on structured metadata information to make recommendations, often using only the star ratings. Such systems ignore the most important information source available in reviews: the textual content. In this work, we apply our text analysis to a recommendation scenario and show that the more detailed textual information can improve rating prediction quality. Our work addresses categorization and sentiment analysis *at the sentence level* as web reviews are typically short and designed to convey detailed information in a few sentences.

The goal of the **URSA** (User Review Structure Analysis) project is to provide a better understanding of user reviewing patterns and to develop tools to better search, understand and access user reviews. In particular, we developed techniques to classify and analyze text- and structure-based web reviews, and used the resulting analysis to improve personalized recommendations for web users.

Our work takes the novel approach of combining natural language processing, machine learning and collaborative filtering to harness the wealth of detailed information available in web reviews.

The remainder of the paper is structured as follows. We report on our sentence classification effort, and on the insight we gained

Twelfth International Workshop on the Web and Databases (WebDB 2009) June 28, 2009, Providence, Rhode Island, USA.

in the process on user-reviewing behavior in Section 2. The classification involved an analysis of the data set to identify predominant categorization topics and sentiments, human annotation of a classifier training set, and testing of the classifier. We propose new recommendation measures that take into account the textual component of user reviews. For this purpose we translated our classified text reviews into text ratings which were used to generate predictions. Our results show that relying on textual information results in better restaurant predictions than using the numerical star ratings given by the users, both in a general recommendation scenario (Section 3) and in a personalized setting (Section 4). We report on related work in Section 5 and conclude in Section 6.

2. STRUCTURE IDENTIFICATION AND ANALYSIS

Web reviews have a combination of linguistic characteristics that depart from the genres traditionally considered in the field of information processing: the language is often quite specific to a particular domain (reviewers of electronic goods, for instance, use many technical terms to describe product features like resolution, battery life, zoom); at the same time reviews are unedited and often contain informal and ungrammatical language. Certain language constructs like sarcasm, make it difficult to identify review sentiment using words as indicators. Finally, reviews often contain anecdotal information, which does not provide useful, or usable, information for the sake of automatic processing.

Our approach to addressing most of the above mentioned challenges is to consider a review not as a unit of text, but as a set of sentences, each with their own topics and sentiments. This added structural information provides valuable information on the textual content at a fine-grain level. We model our approach as a multi-label text classification task for each sentence where labels are both about topics and sentiments. We focused our classification effort on a restaurant review data set, described in Section 2.1. We report on our classification effort in Section 2.2, and on the results of our analysis of user reviewing patterns in Section 2.3 [1].

2.1 Data Set

We extracted our corpus of over 50000 restaurant reviews from Citysearch New York. All reviews present in the system were extracted over the course of one week in 2006.

The corpus contains 5531 restaurants, with associated structured information (location, cuisine type) and a set of reviews. There are 52264 reviews, of which 1359 are editorial reviews and the rest are user reviews. Reviews contain structured metadata (star rating, date) along with text. Typically reviews are small; the average user review has 5.28 sentences. The reviews are written by 32284 distinct users, for whom we only have unique username information.

The data set is sparse: restaurants typically have only a few reviews, with 1388 restaurants having more than 10 reviews; and users typically review few restaurants, with only 299 (non-editorial) users having reviewed more than 10 restaurants.

2.2 Text Review Classification

As the first step of our project, we analyzed the data to identify categories specific to the restaurant reviews domain. These dimensions focus on particular aspects of a restaurant. We identified the following six categories: Food, Service, Price, Ambience, Anecdotes, and Miscellaneous. The first four categories are typical parameters of restaurant reviews (e.g., Zagat ratings). Anecdotal sentences are sentences describing the reviewer's personal experience or context, but that do not usually provide information on the restaurant quality (e.g. *"I knew upon visiting NYC that I wanted to try an orginal deli"*). The Miscellaneous category captures sentences that do not belong to the other five categories and includes sentences that are general recommendations (e.g., *"Your friends will thank you for introducing them to this gem!"*). Sentence categories are not mutually exclusive and overlap is allowed.

Sentence Category	Accuracy	Precision	Recall
FOOD	84.32	81.43	76.72
SERVICE	91.92	81.00	72.94
PRICE	95.52	79.11	73.55
AMBIENCE	90.99	70.10	54.64
ANECDOTES	87.20	49.15	44.26
MISCELLANEOUS	79.40	61.28	64.20
Sentiment	**Accuracy**	**Precision**	**Recall**
POSITIVE	73.32	74.94	76.60
NEGATIVE	79.42	53.23	45.68
NEUTRAL	80.86	32.34	23.54
CONFLICT	92.06	43.96	35.68

Table 1: 7-Fold cross validation of classifier results.

In addition to sentence categories, sentences have an associated sentiment: Positive, Negative, Neutral, or Conflict. Users often seem to compare and contrast good and bad aspects; this mixed sentiment is captured by the Conflict category (e.g., *"The food here is rather good , but only if you like to wait for it"*).

2.2.1 Manual Sentence Annotation

To classify sentences into the above mentioned categories and sentiment classes, we manually annotated a training set of approximately 3400 sentences with both category and sentiment information. To check for agreement, 450 of these sentences were annotated by three different annotators. The kappa coefficient (K) measures pairwise agreement among a set of annotators making category judgments, correcting for expected chance agreement [17]. A Kappa value of 1 implies perfect agreement, the lower the value, the lower the agreement. The inter-annotator agreements for our annotations were very good (Kappa above 0.8) for the Food, Price, and Service categories and Positive sentiment. The Negative sentiment (0.78), Neutral and Conflict sentiments, Miscellaneous and Ambience categories all had good agreements (above 0.6). The ambiguous Anecdotes category is the only one for which the Kappa value was moderate (0.51).

2.2.2 Automatic Sentence Classification

We trained and tested Support Vector Machine classifiers [9] on our manually annotated data (one classifier for each topic and one for each sentiment type). Features for all classifiers were stemmed words (preliminary experiments did not suggest significant improvements in accuracy when more sophisticated features were used for classification). We used svm light[2] with default parameters.

We performed 7-fold cross validation [12] and used accuracy, precision and recall to evaluate the quality of our classification (see Table 1). Precision and recall for the main categories of Food, Service and Price and the Positive sentiment were high (70%), while they were lower for the Anecdotes, Miscellaneous, Neutral and Conflict categories. These low results could be due to the ambiguous nature of these categories but also due to the small amount of training instances in our corpus for these categories in particular.

While the specific categories we identified are tailored for a restaurant scenario, our classification approach could easily be translated to other types of data sets after a topical analysis to identify

[1] `Classified and original data can be downloaded at http://www.dbmi.columbia.edu/noemie/ursa`

[2] `http://svmlight.joachims.org`

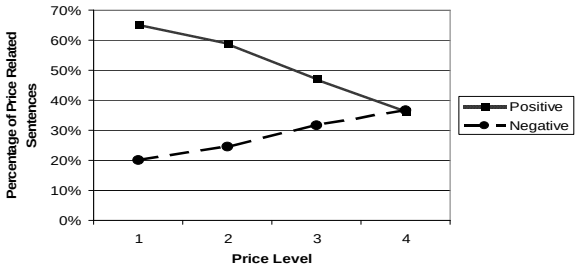

Figure 1: Impact of price level on perception.

product-specific sentence categories.

2.3 Text Review Analysis

To understand trends in reviewing behaviors, we performed an in-depth analysis of the corpus of 52264 user reviews augmented with our automatic classification. Thus, we could study the relation between the textual structure of the reviews and the metadata entered by the reviewers, such as star rating. Due to space limitations we only report on a subset of our findings below.

2.3.1 User Reviewing Trends

Our analysis of the annotated corpus of reviews shows that the sentiment expressed in the reviews was mostly positive (56% of sentences), while only 18% of the review sentences expressed negative sentiment. This is consistent with the star ratings provided by users, with 73% of reviews having a star rating of 4 or 5.

Most reviews describe the food served by the restaurant (32%), while fewer than 17% of the sentences are about the service, 10% are about ambience and 6.5% are about price. The category distribution of reviews is dependent on the cuisine type (metadata) of the restaurant. Restaurants serving French and Italian cuisines have many service related sentences (20%). In contrast, reviews of Chinese restaurants, Delis and Pizzerias focus mostly on Food.

Coarse price level metadata information (numerical value from 1 to 4, 1 being the cheapest) is associated with restaurants in the data set. Figure 1 shows that the number of positive price related sentences decreases and the number of negative price related sentences increases as the price level increases implying, unsurprisingly, that users complain more about prices of expensive restaurant.

2.3.2 Comparing Star Rating with Sentiment

Probably the most important metadata information in reviews is the user-input star rating (from 1 to 5 in our data set, 5 being the highest). We compare this star rating with the sentiment annotation produced by our classifier using the Pearson correlation coefficient [16]. The coefficient ranges from -1 to 1, with -1 for negative correlation, 1 for positive correlation and 0 for no correlation. Our results show a positive correlation (0.45) between the star rating and the percentage of positive sentences in the review, and a negative correlation (-0.48) between the star rating and the percentage of negative sentences. On average, reviews with good ratings of 4 and 5 mainly have positive sentences (71%), and very few negative sentences (6%). In contrast, reviews with bad star ratings (1 or 2) have 5% positive sentences and above 78% negative sentences. These observations and the much finer range of interpretations of text reviews gives us motivation to include text in a restaurant recommendation system, as described in the following section.

3. RATING PREDICTION

Our research hypothesis is that the text of a review (as approximated by its associated topics and sentiments) is a better indicator of the sentiment of the review than the coarse star rating. We test this hypothesis on a recommendation system scenario, and explore whether textually-derived ratings are better predictors than numerical star ratings given a user's restaurant preferences.

Predicting Star Ratings	**TEST I**	**TEST II**
Star rating	1.217	1.295
Sentiment-based text rating	1.098	1.27
Predicting Text Sentiment Ratings	**TEST I**	**TEST II**
Star rating	1.430	1.342
Sentiment-based text rating	1.277	1.374

Table 2: Prediction MSE using the restaurant average for prediction.

We propose two alternative ratings that incorporate text-based information: a rating which relies on sentiment information only (Section 3.2) and one that incorporates topics and sentiment into a regression-based rating (Section 3.3). The later is motivated by the hypothesis that different topics have a different importance in a recommendation scenario as also shown in [10]. We also experiment with different prediction strategies.

3.1 Evaluation Setting

We performed our rating prediction experiments over our restaurant review data set described in Section 2.1. To evaluate the predictive value of our two alternative ratings, we extracted two non-overlapping test sets of around 260 reviews each from the restaurant data set; Test set I contains one review each from users who have rated at least 12 restaurants, Test set II from users who have rated at least 5 restaurants. Therefore, there is less usable user-specific information in Test set II than in Test set I.

We use the popular Mean Squared Error (MSE) accuracy metric to evaluate our prediction techniques [7].

3.2 Sentiment-Based Text Rating

Intuitively, the sentiment expressed in the textual review would be the best indicator of a user's likes and dislikes. To leverage the sentiment information into a single score for each review, we translated our annotated text reviews into a text rating score that can easily be compared to the metadata star rating.

We based the computation of the text rating on the number of Positive and Negative sentences in the review. We ignored Conflict and Neutral sentences as they do not provide insightful information on the quality of the restaurant. For the time being, we ignore the sentence topic information too. We used the formula:

$$TextRating = \left[\frac{P}{P+N} * 4\right] + 1 \quad (1)$$

where P is the number of Positive sentences in the review, and N is the number of Negative sentences. The rating is scaled in the [1:5] range to be comparable to the metadata star rating.

This rating indicates the overall sentiment expressed in the sentences and the review as a whole. We experimented with slightly varying formulas where the importance of P and N sentences was varied, without observing any improvement in rating predictions. We evaluate the robustness of the above mentioned rating with three prediction strategies: one that assumes no additional information is available (section 3.2.1), some metadata information is available (section 3.2.2), or some topical information is available, as derived from our automatic classification (section 3.2.3).

3.2.1 Average Review-based Prediction

In this prediction strategy, we assume that the only information we can rely on is the sentiment (or star) rating. Given a rating (star

Predicting Star Ratings	**TEST I**	**TEST II**
Star rating	1.030	1.117
Sentiment-based text rating	1.051	1.135
Predicting Text Sentiment Ratings	**TEST I**	**TEST II**
Star rating	1.245	1.233
Sentiment-based text rating	1.275	1.199

Table 3: Prediction MSE using cuisine average for prediction.

or sentiment), the rating of a test review is predicted as the average rating of all the other reviews for that particular restaurant.

The resulting MSE values are shown in the top portion of Table 2. The sentiment rating always provides better predicting accuracy, as exhibited by lower MSE values, than the star rating. We computed the statistical significance of this difference using the one-tailed Wilcoxon test [17]. The results for Test set I are significant (p=0.02), but not for Test set II (p=0.12).

Predicting sentiment rating (bottom portion of Table 2) has similar behavior, but provides worse results. We explain this by the fact that text information is ultimately more diverse than star rating.

3.2.2 Average Metadata-based Prediction

We now turn our focus on the impact of available metadata information on prediction accuracy. Our prediction strategy in this case is the following: we predict the rating of a test review for a particular restaurant as the average rating of all restaurants in our corpus that share *the same value for the metadata Cuisine* (we experimented with aggregation over other metadata fields like restaurant location and price-level which we do not report here due to space limitations). Results are shown in Table 3. Using cuisine information improves prediction accuracy. Interestingly, the impact of text-based predictions is not as noticeable in this setting.

Note that this prediction is even more generic than the previous prediction strategy, as it relies on all restaurants belonging to the same cuisine. This can explain why there is less variance between the predictions relying on sentiment-based ratings and on the ones relying on star ratings.

3.2.3 Average Topic-based Prediction

For this prediction strategy, we take advantage of category information as derived from our classification. Our aim is to check whether some categories are more useful for predictions than others. In the two previous prediction strategies, all Positive and Negative sentences in a review were used to compute the text-based rating. In this strategy, we only use the Positive and Negative sentences *belonging to particular categories*. Table 4 shows the results for various category settings for predicting star ratings. The top portion of the table only considered one category. Unsurprisingly, this setting does not perform as well as considering all sentences in the review (Table 2). However, the Food and Miscellaneous categories provide the best results: most sentences are about Food and Food-related information is most meaningful in a restaurant domain; Miscellaneous sentences include general recommendations, which also carry important information for predictions.

The bottom portion of Table 4 shows the prediction results when all but one sentence category were considered for prediction. Excluding the categories that did worse in the single category setting can actually increase the overall prediction quality, as shown by a decrease in the MSE value compared to the default case of Table 2 where all categories are considered (bold values indicate lower MSE values than the corresponding MSE in the default case). This suggest that sentiment information should be used in conjunction with category information when using textual reviews.

Predicting Star Ratings	**TEST I**	**TEST II**
Food	1.215	1.308
Price	1.377	1.424
Service	1.531	1.623
Ambience	1.427	1.559
Anecdotes	1.57	1.676
Miscellaneous	1.221	1.436
All but Food	1.130	1.281
All but Price	**1.096**	1.279
All but Service	**1.096**	**1.269**
All but Ambience	1.115	**1.264**
All but Anecdotes	**1.096**	**1.254**
All but Miscellaneous	1.181	1.352

Table 4: Prediction MSE using the restaurant average for prediction, only considering some categories for the text ratings.

Although the above sentiment-based text rating computation approach shows promising results to improve prediction accuracy, it does not consider category information and with varying degrees of importance. We further investigate the use of categories and sentiments as the basis for recommendation prediction by assigning a regression-based text rating to the reviews, as described below.

3.3 Regression-based Text Rating

We now report on the prediction accuracy when the text rating is derived using multivariate regression. We describe our regression methodology in Section 3.3.1 and present regression-based recommendation results in Sections 3.3.2.

3.3.1 Regression Method

Regression allows us to learn weights to be associated with each sentence type to account for varying importance of the sentence topics. The crucial point is that these weights are learned from the dataset itself, and therefore closely represent how people write reviews in a domain. Our multivariate regression models the user-provided star rating as the dependent variable; the sentence types, represented as *(category,sentiment)* pairs are the independent variables. The value of a sentence type variable for a review is the percentage of sentences of that type in the review.

We use the MATLAB regression function (*mvregress*), which computes the multivariate normal linear regression, to provide estimates using our training data. We performed experiments with different sentence type settings. Our observations show that Neutral and Conflict sentiments do not add significant information and we ignore these. Thus, we use a setting that uses the combination of our two classification sentiments and six classification categories.

The resulting regression weights are shown in Table 6. As expected, the weights confirm the observations from Table 4 in Section 3.2.3: Food and Miscellaneous are the categories that have the highest impact on the perception of a restaurant, while Service is the less important category. Surprisingly, Negative sentences do not negatively impact the score of a review (the corresponding weight is positive) suggesting that even negative information is better than no information at all. This could also be due to the fact that the vast majority of sentences have positive or neutral sentiment, whereas only few sentences have negative information (Section 2.3).

Using the regression weights to compute text rating scores can result in scores that lay outside of the [1:5] range. This makes comparison with star ratings difficult; we scale the text ratings to have the same mean and standard deviation value as the star ratings. In this section we will refer to the unscaled regression-based textual scores as the **raw** scores, in contrast to the **scaled** scores.

Predicting Star Ratings	**TEST I**	**TEST II**
Star rating	1.217	1.295
Regression-based text rating (scaled)	1.089	1.231
Predicting Regression-based Text Ratings (scaled)	**TEST I**	**TEST II**
Star rating	2.680	2.461
Regression-based text rating (scaled)	2.593	2.414
Predicting Regression-based Text Ratings (raw)	**TEST I**	**TEST II**
Regression-based text rating (raw)	0.702	0.742

Table 5: Prediction MSE using the restaurant average for prediction, two-sentiment regression.

Regression Weights	Positive	Negative
Food	4.86	1.53
Price	1.67	1.59
Service	2.61	0.51
Ambience	2.35	2.43
Anecdotes	3.65	2.02
Miscellaneous	5.17	2.27

Table 6: Two-sentiment regression weights.

3.3.2 Two-Sentiment Regression

Table 5 reports on the MSE value for predicting both the metadata star rating and the regression-based (scaled) text rating. Note that the first data row of Table 5 is identical to the first row of Table 2 as these both show the accuracy of using the star metadata information for predicting the star ratings of restaurant-review pairs from the test sets. A direct comparison with Table 2 shows that the regression-based scoring technique performance is comparable to the ad-hoc sentiment scoring technique to predict star ratings. Predicting the regression-based text ratings proves however more difficult than predicting the sentiment-based text ratings, and results in high MSE values.

For completeness, Table 5 also reports on the MSE of predicting text ratings when using the raw text scores. We omitted the comparison with star ratings as the raw scores have a significantly different mean and standard deviation, which skews the predictions and leads to poor results. For text predicting text, the results are significantly better than in the scaled case. However, the standard deviation for raw scores is much smaller than the one for the star rating (and therefore the scaled regression scores) which could mechanically account for part of these low MSE scores.

4. PERSONALIZED RECOMMENDATIONS

A limitation of the prediction metrics used so far is that they are restaurant-based predictions; all users will receive the same prediction for a restaurant regardless of individual preferences. In this section, we investigate personalized recommendation techniques.

We opted to implement a K-Nearest Neighbor algorithm (KNN), a popular collaborative filtering technique [7]. After experimenting with several distance functions, we computed the neighbors using a Pearson distance function with threshold [16] (our implementation uses a threshold value of 5). The threshold is used to account for the number of items in common between users so that users are not considered as very close neighbors on the basis of only one common restaurant rated similarly.

The prediction algorithm uses the average of the K closest neighbors' scores (star rating or scaled text rating) for the target restaurant as the predicted score, if a neighbor has not reviewed the restaurant, it uses the average-case prediction (Section 3.3) for that user.

The resulting MSE values for different values of K are given in Table 7. The corresponding percentage improvements of these values, as a function of K, compared to the restaurant average based MSE of Table 2 are given in Figure 2 for Test set I and Figure 3 for Test set II. For both test sets, the best predictions are around K=10. However, the two test sets perform very differently for low values of K for the regression-based techniques. This can be explained by the fact that users from Test set II have not reviewed many restaurants, therefore their closest ($K < 5$) neighbors might not be good matches as they typically have reviewed few restaurants in common. In contrast, the closest neighbors for the users of Test set I will tend to have very close profiles, which explains the high quality of the corresponding predictions.

Predicting Star Ratings	K	**TEST I**	**TEST II**
Star rating	1	1.210	1.292
	3	1.200	**1.291**
	5	1.194	**1.291**
	10	**1.189**	**1.291**
	20	1.200	1.292
Two-sentiment regression-based text rating (scaled)	1	1.062	1.268
	3	**1.060**	1.235
	5	1.071	1.231
	10	1.066	**1.227**
	20	1.075	1.229

Table 7: Prediction MSE using personalized predictions, two-sentiment regression.

As can be observed, error values are higher, and percentage improvements are lower for Test set II than for Test set I. This can be explained by the fact that the users in Test set II have reviewed (on average) fewer restaurants than the users in Test set I (Section 2.1). Hence, fewer information about these users is available to make accurate predictions (cold start problem).

Getting significant improvements in prediction accuracy is notably difficult. The Netflix Challenge [2] is a real-life example of this problem; Netflix has offered a one-million dollar prize for a technique that would provide a 10% improvement over their in-house Cinematch algorithm. Step-prizes are awarded for each 1% improvement increment. Our MSE percentage improvement of 2.7% (resp. 0.3%) for Test set I (resp. Test set II) can be translated into a rooted mean square error (RMSE) improvement (used by Netflix) of 1.36% (resp. 0.15%). These improvements, while modest for Test set II, show that using text-derived rating for a collaborative filtering techniques is a promising direction. In the future we plan to integrate our text-based scores into more advanced recommendation techniques, in particular we are investigating the use of soft clustering techniques over classified text review data.

5. RELATED WORK

Online reviews are a useful resource for tapping into the vibe of the customers [4]. Accessing and searching text reviews, however, is often frustrating when users only have a vague idea of the product or its features and they need a recommendation. Any large data set requires some filtering based on a user's likes and dislikes. A good survey of the work done in this area and the comparison of several techniques is found in [7] and [3]. Recently, the Netflix challenge [2] has brought a lot of attention to collaborative filtering and recommendation systems. The Netflix data as well as the data used in other projects on recommendation systems like the pioneer GroupLens project [15], consists of highly structured metadata, often only the rating given by a user to a product.

Identifying both topical and sentiment information in the text of a review is an open research question. Review processing has focused on identifying sentiment, product features [5, 13, 6, 18] or a combination of both at once [8, 11, 1, 19]. Hu and Liu [8] and

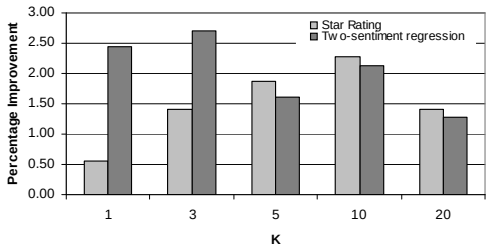

Figure 2: Percentage improvement for KNN over Test I.

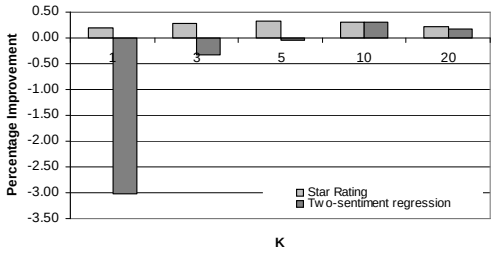

Figure 3: Percentage improvement for KNN over Test II.

other similar studies [14] focus on identifying individual product features and the sentiment expressed towards these features. However, unlike our work these studies do not use the extracted opinions and features in a collaborative filtering scenario. We believe that in a sparse dataset, extracting individual product features results in lesser coverage and a broader topical identification is better for user satisfaction. We approach topical analysis by a sentence level classification. We believe that our classification method, despite the overhead of manually annotating the training set, is better scalable than the popular alternative of Web PMI [14] which requires shooting several queries to search engines.

To the best of our knowledge the only work which incorporates review text analysis in a collaborative filtering system is the recent work by Leung, Chan and Chung [10]. While the authors identify features, they unfortunately do not describe the methods employed for this and do not summarize all their features or roles. Additionally, the evaluation of their recommendation is done by predicting a 2-point or a 3-point rating. We believe that the future generation of recommenders would require finer-grained accurate rating predictions. Our work addresses this need by aiming to predict a 5-point rating scale, commonly used in popular online reviewing systems.

Interestingly, most of the work in sentiment analysis operates at the review level. Our processing unit is a sentence, so that a review is modeled as a combination of topics and sentiments.

In this paper, we present a recommendation algorithm which relies on topic and sentiment information automatically obtained from the text of reviews. We evaluate the performance of our system by making fine grained rating predictions (in the range [1:5]) which is, notably, a harder task than making binary recommendations. This is a novel work that incorporates textual information into a fine grained recommendation system. Additionally, no previous work incorporates the metadata information along with the review text to guide recommendations as we describe in 3.2.2.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the user reviews classification and analysis effort performed as part of our **URSA** project. Our main contribution is the assessment of the impact of text-derived information in predicting the rating of a review in a recommendation system. We show that both topic and sentiment information at the sentence level are useful information to leverage in a review. To the best of our knowledge, this is the first time that the textual component of the review has been considered in such systems, and that user reviews are analyzed and classified at the sentence level.

We are investigating additional refinements to our text-based recommendation, including better text classification strategies, allowing users to get recommendations on specific aspect of restaurants such as food or ambience, and soft clustering-based approaches that group users based on their reviewing styles and interest similarities. In addition, we are interested in the impact of text classification on search over reviews and are implementing tools that allow users to search reviews using category and sentiment information.

We believe this work is paving the way for a better understanding of user reviews and opens interesting future research directions.

7. ACKNOWLEDGMENT

This work was partially supported by a Google Research Award.

8. REFERENCES

[1] N. Archak, A. Ghose, and P. G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *SIGKDD*, 2007.

[2] J. Bennett and S. Lanning. The netflix prize. In *KDD Cup and Workshop*, 2007.

[3] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. pages 43–52, 1998.

[4] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, August 2006.

[5] K. Dave. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *International Conference on World Wide Web*, pages 519–528, 2003.

[6] M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *COLING*, pages 841–847, 2005.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004.

[8] M. Hu and B. Liu. Mining and summarizing customer reviews. In *SIGKDD*, pages 168–177, 2004.

[9] T. Joachims. A support vector method for multivariate performance measures. In *ICML*, 2005.

[10] C. W. ki Leung, S. C. fai Chan, and F. lai Chung. Integrating collaborative filtering and sentiment analysis: A rating inference approach. In *ECAI-Workshop on Recommender Systems*, pages 62–66, 2006.

[11] S.-M. Kim and E. Hovy. Identifying and analyzing judgment opinions. In *HLT-NAACL*, 2006.

[12] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of the International Joint Conference on AI*, 1995.

[13] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *ACL-EMNLP*, pages 79–86, 2002.

[14] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT-EMNLP*, pages 339–346, 2005.

[15] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. pages 175–186, 1994.

[16] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, February 1988.

[17] S. Siegel and J. N. John Castellan. *Nonparametric Statistics for the Behavioral Sciences, Second Edition*. McGraw-Hill, 1988.

[18] B. Snyder and R. Barzilay. Multiple aspect ranking using the good grief algorithm. In *NAACL*, 2007.

[19] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, 2008.