Rutgers DCS Technical Report No. 668
April 2010

URSA - User Review Structure Analysis: Understanding Online Reviewing Trends

by

Gayatree Ganu
Dept. of Computer Science
Rutgers University
Piscataway, New Jersey 08854

Amlie Marian
Dept. of Computer Science
Rutgers University
Piscataway, New Jersey 08854

Nomie Elhadad
Dept. of Computer Science
Columbia University
New York, NY 10032

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 computers to understand, analyze, and aggregate. One consequence of this lack of structure is that searching text reviews is often frustrating for users; keyword searches typically do not provide good results as the same keywords routinely appear in good and in bad reviews. The textual body of reviews contains very rich information and user experience in accessing reviews would be greatly improved if the structure and sentiment information conveyed in the content of the reviews were taken into account. Our work focuses on an analysis of free-text reviews by means of classification of reviews at the sentence level, with respect to both the topic and the sentiment expressed in the sentences. Additionally in this article, we report on the insight on user-reviewing behavior and trends that we gained during our analysis. Our work shows that there is large amount of significant data in the hitherto untapped textual part of user reviews. Our large open-source corpus, of peer-authored text with structure and sentiment information, is a valuable resource for researchers to explore several techniques that have so far relied on structured non-textual data.

1 Introduction

The recent Web 2.0 user-generated content revolution has enabled people to broadcast their knowledge and experience to the masses. Online user reviews are one example of such phenomenon. Products are routinely reviewed and rated by customers on e-commerce destinations such as `amazon.com` and review-dedicated websites like `citysearch.com` and `tripadvisor.com`. Web users, for their part, have whole-heartedly incorporated peer-authored posts into their lives, whether to make purchasing decisions based on peer recommendations or to plan a night out using restaurant and movie reviews. According to a recent survey, online reviews are second only to word of mouth in purchasing influence [1]. Another study [2] shows that 86% of polled individuals find customer reviews extremely or very important. Furthermore, 64% of the individuals report researching products online often, no matter where they buy the product (Web, catalog, store, etc.).

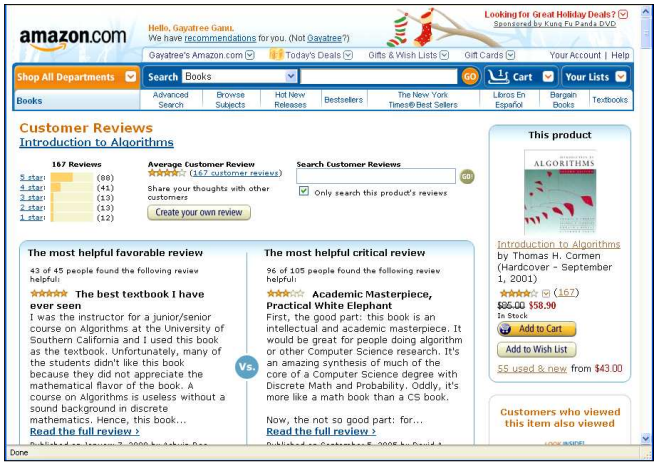

Figure 1: Aggregation technique employed by `amazon.com` to help users access and understand reviews easily.

If online reviews are a trusted and useful source of information for Web users, tools that leverage the valuable information present in reviews are lacking sorely. The sheer number of reviews available for a given product can be overwhelming for users trying to get a comprehensive view of reviewers' opinions. Furthermore, it is often impossible for users to know in advance which reviews contain information relevant to their specific information needs unless they skim all the reviews. Not surprisingly, 78% of polled individuals indicate they spend more than 10 minutes reading reviews for a given product type [2]. Popular websites have started to deploy techniques to aggregate the vast information available in user reviews and to identify reviews with high information content. `amazon.com`, for instance, relies on the star ratings to aggregate reviews and user votes to determine which reviews are helpful (Figure 1). The Internet Movie Database (`imdb.com`) uses the reviewer profile and demographics (Figure 2). Websites dedicated to particular products, like `citysearch.com` for restaurants, ask the reviewers a set of descriptive questions that can be used later as metadata informa-

tion (Crowded, Trendy) when searching products (Figure 3). All these techniques ultimately depend on how much users are willing to contribute, either the reviewers themselves in rating products and answering descriptive questions or the review readers in rating the usefulness of a review. Furthermore, pre-determined metadata are not always flexible enough to represent all the information a user can contribute in a review. Unfortunately, most aggregation techniques so far, ignore the information conveyed in the text of a review.

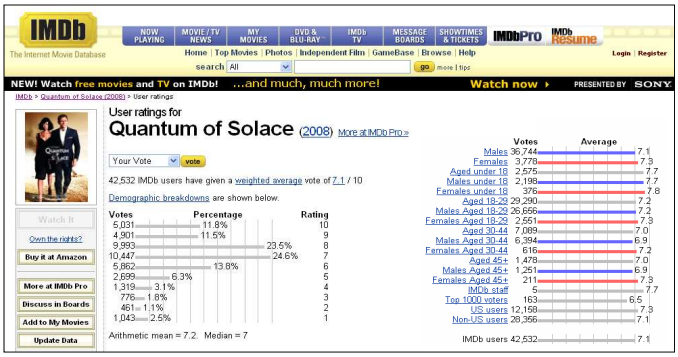

Figure 2: IMDB uses reviewer information to provide aggregated review information.

Figure 3: Citysearch asks the reviewer several optional questions to try and gain information and opinions otherwise available only in text.

1.1 Challenges in Processing Reviews Automatically

User experience would be greatly improved if the structure of the review texts were taken into account, i.e., if the parts of the reviews pertaining to different features of a product (e.g.

food, atmosphere, price, service for a restaurant) were identified, as well as the sentiment of the reviewer towards each feature (e.g.positive, negative or neutral). This information, coupled with the metadata associated with a product (e.g. location and type of cuisine for restaurants), could then be used to analyze, aggregate and search reviews. Consider the scenario of a user looking for products with a set of information needs (i.e., particular features of the product that are important to her). In this scenario, there are several challenges in creating automatic aggregation and search tools that leverage the textual part of reviews.

First, the same individual words routinely appear in good and in bad reviews [3]. As such, a basic keyword-based search engine might not help users identify products with good reviews according to their information needs. Consider the following example:

EXAMPLE 1. *The New York restaurant Bandol in* Citysearch *has several reviews discussing their desserts. However, these reviews are often positive and often negative, as shown here:*

- *"Tiny dessert was $8.00...just plain overpriced for what it is."*
- *"The mussels were fantastic and so was the dessert ...definitely going to be back very soon."*

Another challenge 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 2. *The New York restaurant Lucky Cheng's has 65 user reviews. Out of these, 40 reviews have a 4 or 5 star rating (out of 5 possible stars). Most of the positive reviews, however, focus on and praise the atmosphere 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"*
- *"Dont go for the food because the food is mediocore."*
- *"bottom line they make you feel good when you are in that establishment and that is just as important as the meal itself."*
- *"The food was okay, not great, not bad.[...]Our favorite part, though, was the show!"*

The negative reviews, on the other hand, complain at length about the price and the service. A user not interested in atmosphere would not be interested in this restaurant.

Identifying which features of a product were discussed in a review automatically is a difficult task. Reviewers are creative in their writing, as shown in Example 3.

EXAMPLE 3. *The following sentences are all about the atmosphere and decor of restaurants, even though they share little in common, both in their content and style:*

- *A reviewer writes about the restaurant Nadaman Hakubai: "Unflattering fluorescent lighting and less-than-luxurious furnishings make the space feel more fast food than fine cuisine [...]."*
- *A reviewer about the restaurant Tao: "Great music, beautiful people, great service..... except the part where you don't get seated right away even WITH a reservation."*
- *A reviewer about the restaurant Sea: "Interior designers will be delighted."*

Identifying the sentiment of a statement automatically is an open research question. The same adjective can indicate a positive or a negative sentiment depending on which feature of a product is discussed, as in the following example:

EXAMPLE 4. *The word "cheap" is polysemous in the restaurant domain.*

- *A satisfied reviewer about the restaurant Big Wong: "Cheap eats at a great price!"*
- *A dissatisfied reviewer about the restaurant Chow Bar: "The decor was cheap looking and the service was so-so."*

To complicate matters, some language constructs like sarcasm can confuse any automatic sentiment analysis tool, as in the following example:

EXAMPLE 5. *The New York restaurant Pink Pony has 18 user reviews with an average star rating of 3, indicating that some reviews were positive while some were negative. This makes it further difficult to determine the sentiment of a sarcastic review, as shown here:*

- *"I had been searching really hard for a restaurant in New York where I could really feel unwanted and ignored and I finally found it! The staff ignored my friends and I the entire time we were there...* ***You guys are awesome!****"*

Finally, processing the genre of user reviews automatically differs from processing more traditional written texts, like news stories. The reviews contain unedited, often informal language. Poor spelling, unorthodox grammar and creative punctuation patterns often introduce mistakes when using tools like parsers that have been trained on news stories.

1.2 Contributions of This Work

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 and access user reviews. We focus on the particular domain of restaurant reviews.

We collected a large corpus of restaurant reviews, along with the associated metadata. The text in the reviews was processed to identify sentences and syntactic chunks. We manually annotated a subset of the corpus for topical and sentiment information, according to a coding schema we created, as described in Section 3 and according to the guidelines of Appendix A.

The manual annotation was leveraged to build an automatic classifier that predicts the topical and sentiment annotation of a given review. We report on the classification effort in Section 4.

We applied our automatic classification to the whole corpus. This allowed us to analyze the relation between metadata and textual information. This also enabled us to better understand and uncover interesting patterns in restaurant reviews (Section 5). We show how the information in the annotated corpora can in turn be leveraged in several applications like improving search and social filtering for personalized recommendations as described in Section 6.

Finally, one important contribution of this work is the corpus of reviews itself, along with its manual annotation. Our annotated corpus is a valuable resource for researchers to use textual data in areas where hitherto only structured metadata was used. To facilitate implementations of novel text-based applications we are making this resource available to the community.

The rest of the paper is structured as follows. In Section 2 we give an overview of previous work on sentiment and topic analysis with a focus on peer-authored posts. We then describe our restaurant review corpus in Section 3 and describe the manual annotation of the gold standard. An overview of the rules used for manual annotation are described in Appendix A. This manually annotated subset of the corpus was then used for automatic classification of the rest of the reviews. We describe the classification method, experimental setup and evaluate classifier performance in Section 4. The addition of topical and sentiment information to free-form text enabled us to analyze the content and better understand trends in user behavior (Section 5). We propose some interesting applications using text in reviews in Section 6 and conclude in Section 7.

2 Related Work

Identifying both topical and sentiment information in the text of a review is an open research question. Recently, researchers have gotten interested in this challenge and have investigated several approaches, ranging from leveraging existing lexical resources like WordNet to unsupervised methods. [11] uses, in addition to labeled examples, unlabeled data and prior lexical information for semi-supervised sentiment classification. In our work, we model the

topic and sentiment identification as a supervised classification task.

Review processing has focused on identifying sentiment, product features [12, 13, 14] or a combination of both at once [15, 3, 16]. Interestingly, most of the work in review processing operates at the review level. In our work, however, our processing unit is a sentence, so that a review is modeled as a combination of topics and sentiments.

Extracting relevant and appropriate features for text classification, and annotating the gold standard data is a challenging task. The issues in manual corpus annotation with sentiment information are discussed in [17]. Our work deals with the complexity of topical classification in addition to sentiment classification. Experiments that employ complex strategies for feature extraction, such as proper name substitution, dependency parsing and negating modifications were conducted in [12]. However, the experimental results indicate that except stemming which improves the unigram baseline the linguistic features inversely hurt the performance. In our work we tried using several linguistic features and noticed no significant improvements, therefore we rely on stemmed unigrams for our approach. However, it is shown in [12] that using upto trigrams helps in improving classification accuracy, the work in [18] claims that n-grams with higher n (upto n=6) always outperforms unigram and bigram approaches. Again, our experiments with higher length n-grams did not outperform our baseline approach of using unigrams as features for classification.

Review processing has gather increased attention in recent years. Research shows that online reviews are a useful resource for tapping into the vibe of the customers ([20]). In [3], the sentiment of descriptive adjectives along with product features are identified from the text in reviews. It is also shown that user reviews have a direct economic impact on businesses. Similarly, [19] extracts opinions in blogs and shows how these opinions can be used to predict product sales accurately. Therefore, not only are reviews important to users while making purchasing decisions they are also important to businesses to track feedback. While the study in [3] identifies individual product features, our techniques classify sentences based on wider topics which are the most descriptive of the data in the domain as described in the following section.

3 Annotated Corpus of Restaurant Reviews

Our hypothesis is that there is a large amount of detailed data in the text of the reviews. To tap into this information we propose a sentence level topical and sentiment analysis. For our experiments and analysis we mined 52624 restaurant reviews. The average length of reviews was over 5 sentences per review and only 97 reviews were lacking in textual body. This shows that the analysis of review text in our corpus is a promising direction to explore. In this section, we describe our web-mined corpus in Section 3.1, propose a text classification task in Section 3.2 and present our methodology and evaluation of manual annotation in Section 3.3.

3.1 Corpus Collection

Our corpus consists of restaurant information and reviews from New York Citysearch[1]. Restaurants contain highly structured meta-information, such as name, location, cuisine type, and price level, most likely determined by the Citysearch editorial staff. There are two types of reviews: editorial written by staff, and user-written contributed by web users. All reviews consist of both structured information (such as star rating, username, and date) and a body containing free-form text.

The corpus was collected over the course of one week in 2006. 17,843 restaurants were mined. We kept the restaurants with at least one review, which resulted in 5,531 restaurants. For these restaurants, there were overall 52,624 reviews. Among them, 1,359 were editorial reviews (written under 115 usernames). The other reviews were written by 32,167 different users (we did not mine any private information of the users, and only have unique username identifiers).

The entities in the corpus follow Zipf's law: restaurants typically have only a few reviews, with 1,388 restaurants having more than 10 reviews, 28 restaurants with more than 100 reviews, and one restaurant with the highest number of reviews (242). Similarly, users typically review few restaurants; there are only 299 (non editorial) users who contributed more than 10 reviews, 13 users with more than 50 reviews, and only one user with more than 100 reviews (102).

The average length of user reviews was 5.28 sentences (standard deviation of 2.78) while editorial reviews were 5.44 sentences on average (stdev = 1.99). 96 user reviews contained zero sentences and the maximum length of user reviews was 49 sentences. A single editorial review was of length zero, and the longest contained 11 sentences.

The schema we decided for representing restaurants and their corresponding reviews is shown in Figure 4. For restaurants, the annotation is directly taken from the collected metadata. For reviews, we also included the collected review-specific metadata in our schema, such as star rating. We describe next the annotation schema for the body of the reviews.

3.2 Annotation Schema

There are two types of information we want to annotate in a review: which features of the product were reviewed by the user (we refer to these as topics) and the user's sentiments about these topics. To capture the fact that users typically review several topics in one review and have various sentiments, we decided on sentences as our unit of annotation and processing.

For the restaurant domain, we identified the following topics: Food, Service, Price, Atmosphere, Anecdotes, and Miscellaneous. The first four topics are typical parameters of restaurant reviews (e.g. Zagat ratings). Anecdote sentences describe the reviewer's personal experience or context, without providing much information about the restaurant reviewed. (e.g. *"I knew upon visiting NYC that I wanted to try an orginal deli";"My boyfriend and*

[1] `http://newyork.citysearch.com`

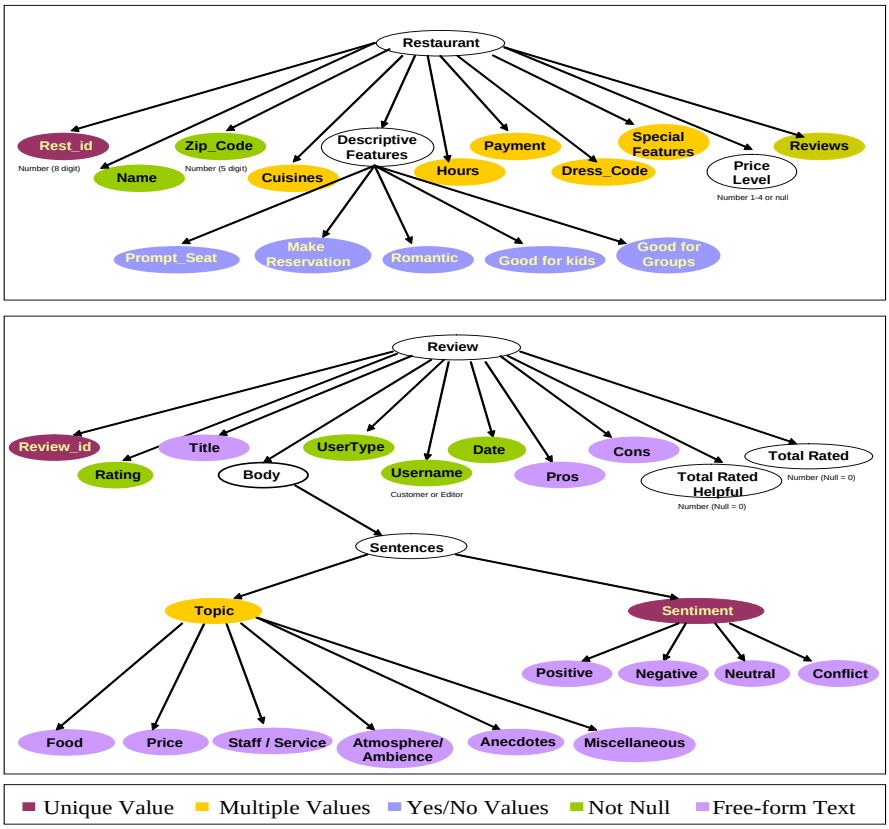

Figure 4: Data Schema for the restaurant review data corpus.

I went there to celebrate my birthday the other night and all I can say is that it was magnificent"). The Miscellaneous topic represents sentences that did not belong to the other five categories and included general recommendations (e.g. *"This restaurant was way overhyped"; "Your friends will thank you for introducing them to this gem!"*). Topics are not mutually exclusive and overlap is allowed. For instance, the sentence: *"The waitress was very patient with us and the food is phenomenal!"* is classified both as Food and Service.

In addition to sentence topics, each sentence has an associated sentiment: Positive, Negative, Neutral, or Conflict. Users often seem to compare and contrast good and bad aspects in the same restaurant; the conflict category is useful to capture sentiment for sentences where various aspects of the restaurant are contrasted (e.g. *"The food here is rather good , but only if you like to wait for it"*). Though we annotate a sentence with possibly more than one category, each sentence of the corpus is assigned exactly one sentiment. We next describe our task of annotating review sentences.

3.3 Manual Annotation

A gold standard annotation of topics and sentiments for 652 reviews (3418 sentences) was created. Reviews were chosen to span several different cuisines, locations and users. Sentences boundaries in the reviews were identified automatically using the OpenNLP toolkit.

The three authors of this paper annotated the gold standard. A screenshot of our Web-based annotation interface is provided in Appendix B. To ensure the manual annotation yields high inter-annotator agreement, annotation guidelines were developed (see Appendix A) and tested through several iterations of annotation on a subset of 86 reviews (450 sentences). All three authors annotated the same 450 sentences to compute inter-annotator agreement. Once the annotation reached acceptable agreement, each author annotated a third of the remaining reviews.

Cohen's Kappa was used to compute the inter-annotator agreement [5]. A Kappa value of 1 implies perfect agreement, the lower the value, the lower the agreement. Table 1 shows the average Kappa values for each topic and sentiment. The Food, Price, and Service topics and the Positive sentiment had almost perfect agreement (K>0.8). The Negative sentiment (K=0.78), Neutral and Conflict sentiments, Miscellaneous and Ambience topics all had substantial agreements (K>0.6). The ambiguous Anecdotes category is the only one for which the Kappa value was moderate (0.51).

The reviews in the gold standard were chosen to be representative of the overall corpus in terms of cuisine and star ratings (67 reviews with 1 star, 42 with 2 stars, 56 with 3 stars, 145 with 4 stars and 342 with 5 stars). The manual annotation indicates that the topics and sentiments are unbalanced (see Table 2). Users, not surprisingly, focus on reviewing the food (38.39% of gold standard sentences had a FOOD label). The majority of sentences are positive (54.45% of sentences have a POSITIVE label).

Topic	Kappa
FOOD	0.88
PRICE	0.87
STAFF	0.84
AMBIENCE	0.62
ANECDOTES	0.51
MISCELLANEOUS	0.60

Sentiment	Kappa
POSITIVE	0.83
NEGATIVE	0.78
NEUTRAL	0.65
CONFLICT	0.60

Table 1: Inter-annotator agreement for the gold standard.

Topic	Proportion
FOOD	38.39%
PRICE	9.77%
STAFF	18.29%
AMBIENCE	13.02%
ANECDOTES	12.7%
MISCELLANEOUS	26.92%

Sentiment	Proportion
POSITIVE	54.45%
NEGATIVE	21.81%
NEUTRAL	17.45%
CONFLICT	6.29%

Table 2: Proportion of topics and sentiment sentences in the gold standard (A sentence can have only one sentiment but several topics).

4 Automatic Classification of Topics and Sentiments

We used the representative hand-annotated gold standard for training and testing automatic classifiers for identify topics and sentiments in the rest of the corpus. The details are in the following section.

4.1 Classification Method and Features

Since a given sentence can have more than one topic assigned to it, the topic classification is a multilabel classification problem. There are two standard approaches to multilabel classification [9]: binary relevance and label powerset. A binary relevance classifier trains one binary classifier for each category (each instance is categorized as having this label or not). In our case, this translates in training six independent classifiers, one for each topic. This approach ignores the possible correlations among labels. A label powerset multilabel classifier instead builds a classifier for each possible label subsets. In our case, this translates in training 2^6 independent binary classifiers. Furthermore, the resulting classifiers have very sparse training sets, since not all combinations of topics are present in the training data. Since our gold standard is not large, this caveat becomes problematic. Thus, we opted for a binary relevance framework. A similar framework was used for the task of sentiment classification. For both topical and sentiment classification, we relied on support vector machines.

We experimented with lexical features (unigrams and bigrams). For the sentiment clas-

Topic	Acc.	P	R
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	Acc.	P	R
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 3: Classification results (accuracy, precision, and recall) with 7-fold cross validation.

sification, we also experimented with selecting only words with particular part-of-speech (adjectives and adverbs for the sentiment analysis for instance), syntactic features based on dependency parses, and semantic features (relying on WordNet to create sets of adjectives with different polarities). Because we want our method to stay domain independent, we did not investigate features specific to the restaurant domain. Similar to our approach, the work in [13] shows that SVMs using stemmed unigrams as features demonstrate very good classifier performance.

4.2 Experimental Setup and Results

The six topic classifiers and the sentiment classifier were built using svm light. We performed 7-fold cross validation [6] and relied on accuracy, precision and recall to evaluate the quality of our classification.

We experimented with different feature combinations, but all performed with little statistical difference from a baseline approach which relies on stemmed unigrams only. For simplicity and speed, we adopt this baseline approach for all further analysis (SVMs using stemmed unigrams as features). Table 3 shows the classification performance for the baseline topic and sentiment classifiers.

Most categories are classified with high precision. Precision and recall for the most frequent categories of Food, Service, Price, and Positive sentiment were high (greater than 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 (these are the same categories for which inter-annotator was only moderate), but also due to the small amount of training instances in our corpus for these particular categories.

Another potential way to interpret the results is that, like in Example 4 presented earlier, some words mean different things in different contexts and could confuse both the topic and sentiment classification. To verify this hypothesis we conducted the following experiment: we controlled for the cuisine type in both the training and testing sets. Sentences of a particular cuisine were used to train a classifier, and the classifier was tested with sentences belonging to the same cuisine. These classifiers yield significantly more accurate results than in the general case (see Figure 5). This result confirms our intuition. Because the gold standard does not span all the cuisines in the dataset with sufficient representative sentences (10 most popular categories in the corpus were included in the gold standard, and the results in

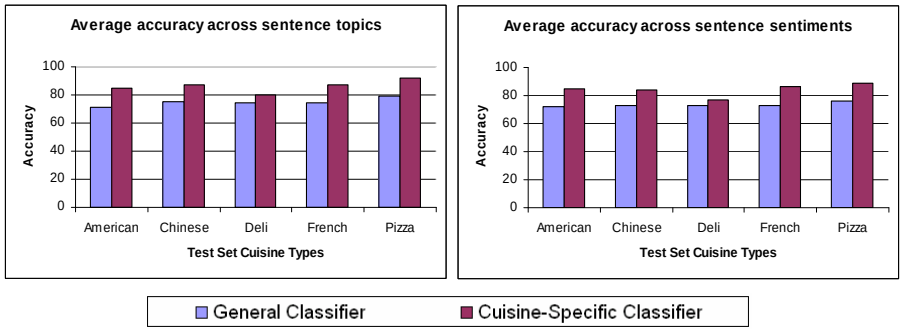

Figure 5: Comparison between general and cuisine specific classification.

Figure 5 are observed using 5 cuisine types with an average of 268 sentences of each cuisine type), we do not investigate this type of stratification further. We acknowledge, however, the potential improvement in classification accuracy.

5 User Reviewing Trends

To understand trends in reviewing behaviors, we performed an in-depth analysis of the entire corpus of user reviews. We applied our automatic classification to all the reviews. We investigated the relations between the textual content of the reviews and the metadata entered by the reviewers, as described in this section. While we uncovered some surprising and interesting trends, we also confirmed some obvious and expected behavior. The later shows that our classification is sound and our analysis of textual information conforms to the expected behaviors in user reviews. To the best of our knowledge, this is the first study of its kind that does an in-depth study of the textual content in user reviews.

5.1 Category Distribution

Our first step was to develop an understanding of the distribution of sentences in terms of topics and sentiments. Most reviews have sentences focusing on the food served by the restaurant (32%), while fewer than 16% of the sentences are about the service, 11% are about ambience and 6.5% are about price.

The distribution of sentence categories and sentiments are shown in Figure 6. The sentence category and sentiment is also affected by the type of the restaurant, defined by the metadata, as shown in the following sections.

5.2 Sentiment Distribution

Our analysis of the annotated corpus of 52,264 user reviews shows that the sentiment expressed in the reviews was mostly positive (56% of sentences), while only fewer than 18% of the sentences expressed negative sentiment (Figure 6). This is consistent with the star ratings information provided by users, with 73% of reviews having a star rating of 4 or 5.

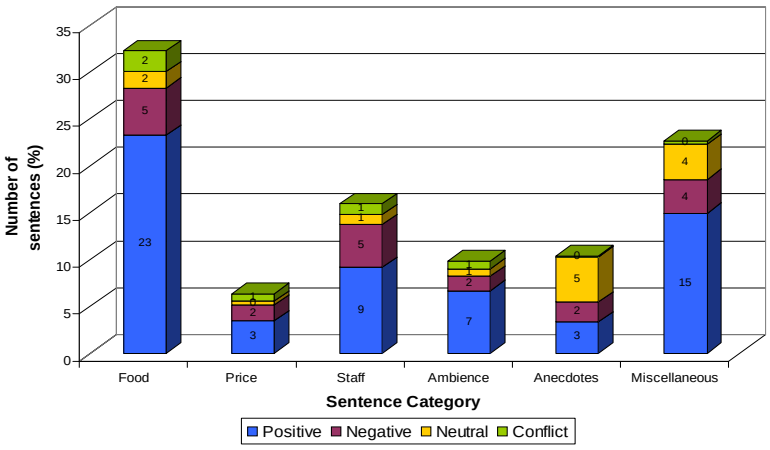

Figure 6: Distribution of the categories and sentiments of automatically classified sentences.

5.3 Dependence on Cuisine Type

Our observation in Section 4.2 led us to the intuition that in addition to the free-form textual data in reviews there is invaluable information in the structured metadata associated with the reviews, and in fact these two parts of the reviews are often related. This led us to explore the correlation between the unstructured text and the structured metadata.

The category distribution of reviews is dependent on the cuisine classification (metadata) of the restaurant. Restaurants serving French and Italian cuisines have many Service related sentences (20%). Surprisingly most of these sentences for French restaurants were Negative (50%) while for Italian restaurants these sentences were mostly Positive (72%). In contrast, reviews of Chinese restaurants, Delis and Pizzerias focus mostly on Food.

5.4 Relation with Location

Location influences the distribution of sentence categories and sentiment in reviews. All restaurants have an associated address containing a zip code which we mapped to the five Boros of New York - Bronx, Brooklyn, Manhattan, Queens and Staten Island. This gives us coarse geographical information, and we realize that due to the diversity of localities, say within Manhattan itself, a finer grained analysis would be interesting. Nevertheless, we still uncover some interesting trends.

Restaurants in Manhattan have more negative reviews (greater than 18%) than other restaurants (13% on average). Reviews for restaurants in Manhattan also have a higher focus on the Service and the Ambience, in contrast to reviews of restaurants in Bronx and Brooklyn which mainly focus on Food.

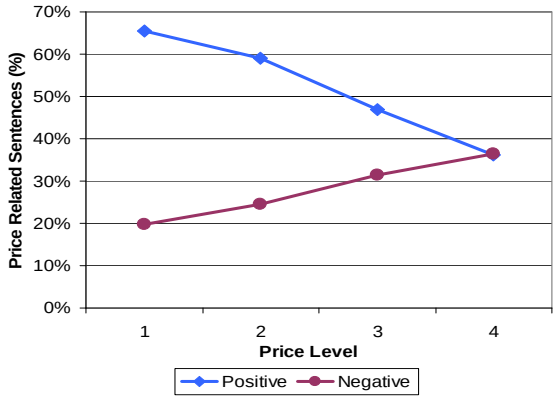

Figure 7: Effect of price level of restaurant on review sentiment.

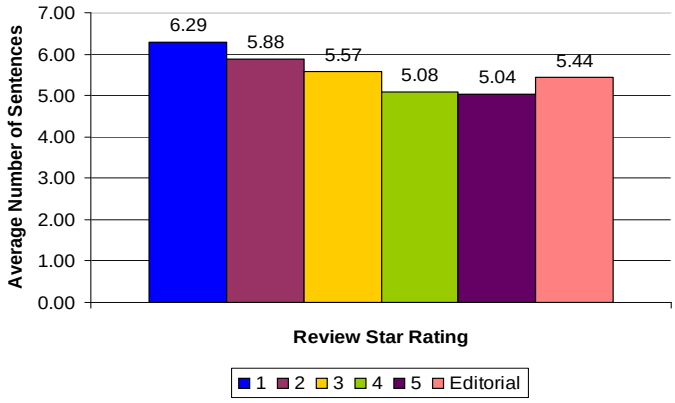

Figure 8: Relation between the length of sentences and the rating of reviews.

5.5 Dependence on Price-Level

Coarse price level metadata information (from 1 to 4, 1 being the cheapest) is associated with restaurants in the data set. Figure 7 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.

5.6 Length of Reviews

The length of the review also depends on the sentiment associated with the reviews: reviews with a bad star rating of 1 were longer (6.3 sentences on average) than reviews with good star ratings of 5, as shown in Figure 8. Most editorial reviews do not have an associated numerical star rating. Interestingly, the average length of the rated editorial reviews is 6.56 sentences while unrated editorial reviews are on average 3.78 sentences long.

5.7 Editorial Reviews and User Reviews

Among the 5531 restaurants in the corpus, 1359 restaurants had editorial reviews. Editorial reviews were significantly different than user reviews. This difference is shown in Figure 9, where we observed that editorial reviews clearly focus on the Food and Ambience of a restaurant while user reviews have many sentences which are about the Price, Service and Anecdotal information; users tend to evaluate restaurants with a different perspective than the editorial staff. One would expect that editorial reviews would be neutral in sentiment, or would at least have equal positive and negative sentences. However, our corpus shows that over 67% sentences of editorial reviews were positive, about 18% were negative, while only less than 12% were neutral. These observations led us to believe that most editorial reviews are more appreciative than critical of restaurants.

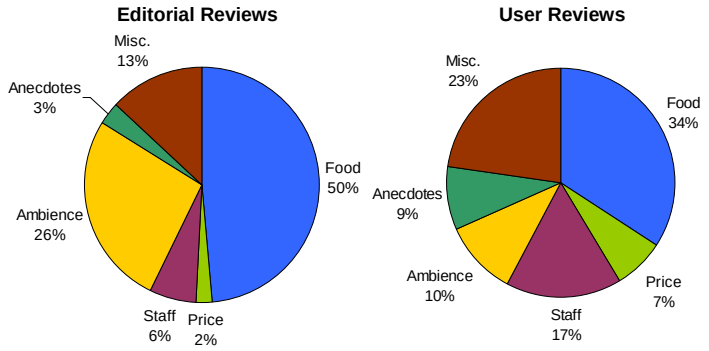

Figure 9: Comparing editorial reviews with user reviews.

5.8 Comparison with Star Rating

Probably the most important metadata information in reviews is the user-inputted star rating (from 1 to 5 in our data set, 5 being the highest). The star rating is often used as the main filter for ordering reviews presented to a customer. In today's reviewing systems, the average star rating is the most important indicator (often the only indicator) used to assess the quality of the product.

We compare this star rating with the sentiment annotation produced by our classifier using the Pearson correlation coefficient [7]. 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. Our observations showed that, on average, restaurants with good ratings of 4 and 5 mainly have positive sentences (69%), and very few negative sentences (9%). In contrast, restaurants with bad star rating (1 or 2) have more than 26% positive sentences and as few as 41% negative sentences. These observations, and the much finer

range of interpretations of text reviews gives us motivation to include text information in determining the quality of the products.

To the best of our knowledge, this is the first study that describes the analysis of trends in user reviewing pattern. We developed valuable insight by conducting this analysis. For instance, in the future we would like to reduce the skew in the distribution of topics in our gold standard and include sufficient representatives of each sentence class to train our classifiers. If we wished to include more staff related sentences in our gold-standard, we now know that we would probably benefit from annotating French and Italian restaurants, especially those in Manhattan.

Therefore, our topic and sentiment information can be used for better accessing and understanding reviews. In the following section we propose some interesting applications that use the augmented text in reviews, and show some preliminary results.

6 Impact

As described throughout the paper, augmenting free-form text with structural information can make it useful in areas that have hitherto used only structured information. There are several applications that can use our classified text, we describe a few of these which are future works in the **URSA** project.

6.1 Text and Structure Search

Rather than using simple keyword searches to access user reviews, we can take advantage of the review's structural information to improve search results. For this purpose we are currently working on developing techniques inspired from text and structure search in XML ([8]). Most of these techniques rely on finding keywords within a specific structural context, and allow for some flexibility in the structure of the data. Specifically, we are working on techniques to allow users to specify some structural constraints at query time but allow for some approximation of both the structure and the content resulting in ranked query results. A query for reviews containing the keyword "amazing" (content) in a food context (structure) would then return those reviews that match the query specifications exactly, but would also consider approximate matches such as reviews with the word "amazing" appearing in a different context, or reviews with words similar to "amazing" appearing in food-related sentences; these approximate matches would be scored according to their similarity to the original query conditions.

6.2 Finding Similar Users

In several applications aimed at using peer-authored data, users are often clustered to make meaningful groups of people having similar tastes. This is particularly necessary in a sparse dataset like ours, where we cannot gather sufficient information about a user as a stand-alone entity. In such cases, there is a lot of information to be gained from a user's neighbors,

found via clustering users. For the purpose of this preliminary analysis, we clustered the 32284 users in the corpus using the popular K-Means clustering with different values for the parameter K. We compared two strategies: first using only metadata like location, zip code and price level for features, and in the second case adding the user's preference as features. The preference in restaurant features and the sentiment expressed towards these features was derived from the textual data using the sentence category and sentiment. Our observation shows that the clustering with textual data creates different clusters which can help in addressing the problem described in Example 2. For example, using the annotated textual data we can now have clusters of users who not only all like expensive, Italian restaurants near Times Square, but like them for their ambience and not necessarily their food. Consider the following example observed from the clustering results:

EXAMPLE 5 *For k=20, without using textual information, a cluster was found having 278 users with reviews for steakhouses in the zip code of 10017 and which were expensive (price level 4). After the inclusion of textual information, a similar cluster (having similar cuisine type, location and price level) with 376 users was found. However, a difference in cluster membership was observed as described below.*

User ***desnr02*** *was originally clustered with user* ***anderson1****, but after adding the sentence information* ***anderson1*** *was no longer in the same cluster. Instead* ***Aisle79*** *was now clustered with* ***desnr02****. We observed that* ***desnr02*** *and* ***Aisle79*** *often write about the staff in a restaurant, but* ***anderson1*** *does not. Additionally,* ***anderson1*** *has negative reviews about food as against the positive emphasis by* ***desnr02****. This difference in review focuses implies that* ***desnr02*** *and* ***anderson1*** *are not very similar and should be placed in different clusters, which was correctly captured by a text-based clustering. Therefore, the above experiment indicates that using text in clustering could prove to be very useful for some applications.*

Our clustering analysis is preliminary, mainly because of the sparse dataset. We are currently trying alternative clustering techniques which have worked well on other sparse datasets.

6.3 Similarity Search in Social Networks

Performing simple searches on user reviews is an important problem. However, allowing for complex similarity searches in a social network setting is a promising direction. Users tend to cluster into groups that share the same likes and dislikes. Being able to find products that similar users like would be a definite added value for user reviewing systems. Similarly, presenting products to users which are similar to other products they have reviewed and liked, rather than asking users to specify individual search parameters, would allow for more intuitive and expressive searches.

Towards this step of the project, we have worked on making personalized recommendations via collaborative filtering. We use the textual data along with the metadata, and our preliminary results show that text is often a better predictor of the star rating assigned by a user for a restaurant.

7 Conclusions

The goal of the **URSA** project is to analyze and ultimately understand the way users review products online. In this paper, we focus on the domain of restaurant reviews, but our methods are domain-independent, once the data is analyzed and the domain-specific topics are identified.

In this article, we describe our corpus of restaurant reviews which we make available to the research community. The corpus is large and contains both metadata and textual content. A gold standard of reviews was manually annotated for topic and sentiment information at the sentence level with high inter-annotator agreement. Relying on standard classification tools and the gold standard, we annotated the entire corpus with topic and sentiment information automatically. This data was analyzed for trends in reviewing. We discuss potential applications for our analysis: text combined with structure search and collaborative filtering. To the best of our knowledge, no other study uses the textual component of the review in such systems.

There are several limitations in our current classification system, which lay the path for our future work. First, in our manual annotation, a sentence can have only one sentiment. To capture sentences that discuss several topics but have different opinions, we have introduced a Conflict category. Ideally, each topic should have its corresponding sentiment; this can be achieved by a phrase level annotation and classification. Secondly, the binary relevance approach to topic classification ignores the potential relations among the different topics. Our trend analysis shows that the metadata, topics and sentiments are often correlated. This simplifying approach might affect negatively the classifiers' performance. Finally, the features we investigated so far did not yield better results in the topic and sentiment classification when compared to a baseline set of stemmed unigrams. In our future work, we plan to further investigate the classification method; to choose a more appropriate multi-label classification framework and to identify better features. At a higher level, we plan to continue our work on incorporating text-based content into tools that help users access and search online reviews.

With the recent growing interest and popularity of peer-authored web posts, there is tremendous amount of valuable data in the form of free-form text. This work takes the novel approach of combining Natural Language Processing, Machine Learning and Information Retrieval techniques to harness the wealth of detailed information available in web user reviews.

References

[1] Marketing Charts, http://www.marketingcharts.com/interactive/online-reviews-second-only-to-word-of-mouth-in-purchase-influence-6968/ (2008)

[2] Power Reviews, http://www.powerreviews.com/social-shopping/news/press_breed_11122007.html (2007)

[3] Nikolay Archak and Anindya Ghose and Panagiotis G. Ipeirotis, Show me the money!: deriving the pricing power of product features by mining consumer reviews, Proc. of SIGKDD (2007)

[4] Thorsten Joachims, A Support Vector Method for Multivariate Performance Measures, Proc. of the 22nd International Conference on Machine Learning (2005)

[5] Siegel, Sidney and N. J. Castellan, Jr., Nonparametric Statistics for the Behavioral Sciences, Second Edition, McGraw-Hill (1988)

[6] Ron Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Proc. of the International Joint Conference on Artificial Intelligence (1995)

[7] J. L. Rodgers and W. A. Nicewander, Thirteen ways to look at the correlation coefficient, The American Statistician, 42, 59-66 (1988)

[8] Sihem Amer-Yahia and Laks V. S. Lakshmanan and Shashank Pandit, FleXPath: Flexible Structure and Full-Text Querying for XML, SIGMOD Conference, 83-94 (2004)

[9] G. Tsoumakas and I. Katakis, Multi-Label Classification: An Overview, International Journal of Data Warehousing and Mining, 3(3):1-13, (2007)

[10] G. Tsoumakas and I. Vlahavas, Random k -Labelsets: An Ensemble Method for Multilabel Classification, Proc. of the European Conference on Machine Learning, 406-417 (2007)

[11] S. Vikas and M. Prem, Document-word Co-regularization for Semi-supervised Sentiment Analysis, ICDM '08: Proc. of the 2008 Eighth IEEE International Conference on Data Mining, 1025-1030 2008

[12] K. Dave, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. Proc. Of the 12th International Conference on World Wide Web, 519-528 2003

[13] B. Pang and L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment Classification using Machine Learning Techniques. Proc. of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 79-86 2002

[14] M. Gamon, Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. Proc. of COLING, 841-847 2005

[15] M.Hu and B. Liu, Mining and Summarizing Customer Reviews. Proc. of SIGKDD, 168-177 2004.

[16] I. Titov and R. McDonald, A Joint Model of Text and Aspect Ratings for Sentiment Summarization, Proc. of the ACL Conference, 2008

[17] W. Janyce and W. Theresa and C. Claire, Annotating Expressions of Opinions and Emotions in Language, Language Resources and Evaluation 2005

[18] H. Cui and V. Mittal and M. Datar, Comparative Experiments on Sentiment Classification for Online Product Reviews, Proc. of National Conference on AI 2006

[19] L. Yang and H. Xiangji and A. Aijun and Y. Xiaohui, ARSA: A Sentiment-Aware Model for Predicting Sales Performance using Blogs, Proc. of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2007

[20] J. Chevalier and D. Mayzlin, The Effect of Word of Mouth on Sales: Online Book Reviews, Journal of Marketing Research, 345-354 2006

A Guidelines for Annotation

After several iterations and group discussions, we decided on the following guidelines for manual annotation.

A.1 General Guidelines

1. A sentence can be annotated in one or more different categories, and exactly one sentiment class.

2. No sentence should be considered in the light of previous sentences. For example, if there was a review which spoke only about the ambience, and the last sentence of the review was "It was great!!" then this sentence should not be classified as Ambience too.

A.2 Categories

Food:

1. Sentences that are about the dishes, specialties, the way the food tastes or is prepared should be annotated as food.

2. Sentences that are about portions only should be in this category. However, if the food is compared to the price the annotation should also include the Price category.

3. Sentences about how the food is/was served or presented should be in this category and should be in the Staff/Service category.

4. Sentences that comment on the menu, wines and drinks are to be annotated as Food. Sentences that should be considered to belong to this category:

5. Examples: “A bakery , serving breakfast pastry , pan ini , cakes , cookies , coffee and wine .”, “Seitan – black pepper and also teriyaki are delicious .”, “The menu is unique without being trendy and the food is fantastic .”

Price:

1. Sentences that directly have dollar amounts or prices for some items on the menu.
2. Sentences that are about portions, value for money and worth.
3. Examples: “You are better off making your own sandwich then spending $ 6.00 for nothing .”, “I would have left if i had n’t already paid $ 28 .”

Staff:

1. Sentences that are about waiting time, mean/rude or polite/helpful waiters.
2. If something is below expectations that is due to negligence of staff, like dirty tables, cutlery.
3. Examples:“The presentation is clean and fresh , and the service was phenomenal .”, “My only complaint concerns the hours of operation .”

Ambience:

1. Sentences about lighting, seating arrangements and generally about the atmosphere.
2. Specialty restaurants or theme restaurants will have sentences in this category.
3. Sentences that talk about the crowd at the restaurant.
4. Examples: “Hope they get outdoor seating soon .”, “Industrial gray dominates the color scheme , but soft lighting warms the space and eclectic dance music (spun by a DJ on weekends) lifts the mood .”

Anecdotes:

1. This category and the Miscellaneous category are often confused and result in low accuracy results.
2. Any sentence that is a story or describes a personal experience belongs here.
3. If there is a sentence which says, “We went on Sunday and the service was horrible”, then it is classified as both Anecdotes and Staff.
4. Personal choice sentences like “I will never go there again”, should also be classified in this category.

5. Examples: “I ’ve gone several times , both with friends and in large groups , and I ’m never disappointed”, “Since I live in downtown Brooklyn , Im always on the lookout for unique , chick , and affordable places to go for an intimate dinner with my boyfriend or a leisure brunch with friends .”, “We had lunch the other day at XXX it was a joke .”

Miscellaneous:

1. A Miscellaneous sentence is any sentence that does not match any of the above five category descriptions.
2. Quite often there are single word sentences like “Great!” or “:)” that will belong to this category.
3. General recommendation sentences should belong to this category.
4. Examples: “I recommend this restaurant to vegetarians and to non-vegetarians .”, “Be nice to yourself and go to XXX .”, “wow , what a find .”, “I ’m totaly going back soon .”

A.3 Sentiments

Positive:

1. Sentences that have only positive words or meaning.
2. Types of sentences included could be good/delicious food, polite/friendly staff, pleasant atmosphere, affordable price. Should also include anecdotes about good experience and even single word sentences like “Great!” or “:)”
3. If a restaurant is positively recommended then it should belong to this category.
4. Examples: “I ca n’t wait to go back to New York to try something else on the menu! !”, “The service was wonderful - very warm and attentive .”

Negative:

1. Sentences that have only negative words or meaning.
2. Examples: “I recommed that you walk right past this place .”, “I have never had an order screwed up so badly .”

Conflict:

1. This sentiment and the Neutral sentiment are problematic with low accuracy.
2. Sentences that have a positive and a negative clause.

3. Even if the majority of the sentence is about the positive aspects, but there are some negative sentiments, the sentence is still a conflict sentence.

4. Examples: “They were very friendly , but food was horriable .”, “Only negative was that it was very loud when we got there (at 7pm) but as dinner progressed it got quieter .”, “The wine list is meager but the unforgettable food makes up for anything missing.”

Neutral:

1. If no sentiment is mentioned the sentence is Neutral.

2. Recommendations will not belong to this category as they are either positive or negative.

3. Many anecdotes will fall in this category like “We went on Sunday”.

4. Examples: “Need to make reservations in advance .”, “This place opened a few weeks ago in my neighborhood .”

B Annotation Interface

The online annotation interface was as follows:

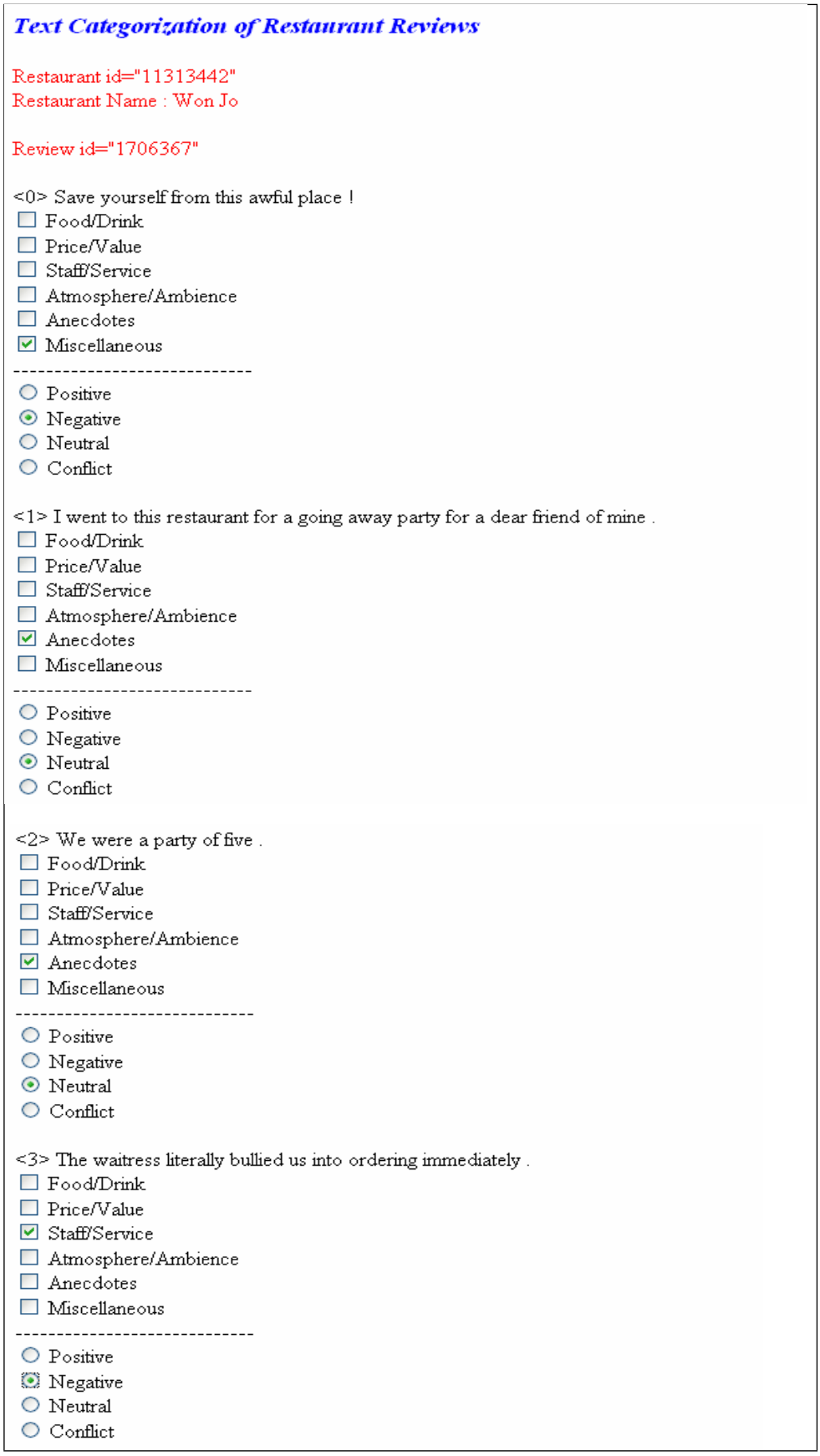

Figure 10: Annotation Interface.