Beyond the Stars: Improving Rating Predictions using Review Text Content

Gayatree Ganu
Computer Science
Rutgers University

Amélie Marian
Computer Science
Rutgers University

Noémie Elhadad
Biomedical Informatics
Columbia University
Project Description

User Review Structure Analysis (URSA)

Aim:
Better understanding of user reviews
Better search and access of user reviews

Tasks:
Structure Identification and Analysis ✔
Text and Structure Search
Similarity Search in Social Networks
Online Reviewing Systems: Amazon
Online Reviewing Systems: Citysearch

Structured Metadata
Online Reviewing Systems: Citysearch

Rate It

<table>
<thead>
<tr>
<th></th>
<th>Not Recommended</th>
<th>Below Average</th>
<th>Average</th>
<th>Recommended</th>
<th>Highly Recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atmosphere</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food and Drinks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pros/Cons

Thumbs up? Thumbs down? List highlights and drawbacks. (Examples: Ambiance, Parking, Service)

Pros: 

Cons: 

Comments

<table>
<thead>
<tr>
<th>Statement</th>
<th>Yes</th>
<th>No</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>The bar was crowded</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The bar was popular and trendy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The bar was romantic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This was a good after-work bar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This bar was a good place to dance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This bar was a good place to meet people</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The bar had a good happy hour</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---

Data in Reviews

Restaurant-specific Metadata

Review Metadata

User inputted free-form text

---
Data in Reviews - Recap

- Highly structured metadata
  - Star Ratings, Bought/Not Bought, Structured data in forms

- Untapped Information
  - Review Text
    - Contains sentiment information
    - Contains information on product specific features
Online Reviews: Uses

➢ For the **User** to decide to buy a product, see a movie, or go to a restaurant.

➢ For the **Business** to track user feedback.
Why do we care?

• **86%** find that customer **reviews are very important**.

• **64%** report **researching products online often, no matter where they buy the product**

[Marketing Charts]

• Users are inconvenienced because:
  - Large number of reviews available
  - Finding relevant reviews
  - Vague or undefined information need

• **78%** indicate they **spend more than 10 minutes reading reviews**

[Power Reviews]
Outline

• Motivations
• **Challenges in Analyzing text**
  • Data Description
  • Text Review Classification
  • Text Based Recommendation System
  • Personalized Rating Prediction
  • Future Directions
• Conclusions
Examples and Challenges in Text

1. Keyword search is ineffective. Often we need more interesting searches like “Find all romantic restaurants”.

2. Sarcasm makes it hard to determine the sentiment.

3. Anecdotes and Miscellaneous information do not add much

4. Adjectives can convey positive or negative sentiment

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 décor was cheap looking and the service was so-so.”
Examples and Challenges in Text

5. Ratings might be reflective of specific product features.

Restaurant Name: **Lucky Cheng’s** (Has 65 user reviews, 40 reviews have a 4 or 5 star rating)

- “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!”

6. Star rating is not a correct indicator of quality

Restaurant Name: **Heartland Brewery Midtown** (Rating 3)

“...they know what they’re doing. I’ve been here a bunch of times now and the service is always outstanding. The food is above average...”

Restaurant Name: **Café Noir** (Rating 3)

“...people who work there are evil and incompetent!! The service was terrible, we had to wait for everything... The whole set up is truly unprofessional...”
Outline

• Motivations
• Challenges in analyzing text
• **Data Description**
  - Augmented Schema
• Text Review Classification
• Text Based Recommendation System
• Personalized Rating Prediction
• Future Directions
• Conclusions
Data Description


- The corpus contains:
  - 5531 restaurants, with
    - associated structured information (name, location, cuisine type) and
    - a set of reviews
  - 52264 reviews, of which 1359 are editorial reviews and the rest are user reviews. Reviews contain
    - structured information (star rating, username, date) along with
    - unstructured text (title, body, pros, cons)
  - 32284 distinct users. We only have unique username information.

- Dataset accessible at http://www.research.rutgers.edu/~ggunu/datasets
Data Description

• Dataset is very sparse:
  - 1388 restaurants have more than 10 reviews.
    28 have more than 100.
  - 327 users have reviewed 10+ restaurants.
    13 users with 50+ reviews
    1 with more than 100 reviews.
• Average length of user reviews was 5.28 sentences.
• 96 user reviews contained zero sentences
Data Schema

Restaurant

- Rest_id (Number (8 digit))
- Zip_Code (Number (5 digit))
- Name
- Cuisines
- Descriptive Features
- Payment
- Hours
- Dress_Code
- Special Features
- Reviews
- Price Level (Number 1-4 or null)
- Prompt_Seat
- Make Reservation
- Romantic
- Good for kids
- Good for Groups

- Unique Value
- Multiple Values
- Yes/No Values
- Not Null
- Free-form Text
Our contribution to the Data Schema

Data Schema

- **Title**
- **Body**
- **Username**
- **Date**
- **Pros**
- **Cons**
- **Total Rated**
- **Total Rated Helpful**

**Review**

- **Review_id**
- **Rating**
- **UserType**
- **Username**
- **Customer or Editor**

**Sentences**

- **Topic**
  - Food
  - Price
  - Staff / Service
  - Atmosphere / Ambience
  - Anecdotes
  - Miscellaneous

- **Sentiment**
  - Positive
  - Negative
  - Neutral
  - Conflict

**Unique Value**

**Multiple Values**

**Yes/No Values**

**Not Null**

**Free-form Text**
Outline

- Motivations
- Challenges in analyzing text
- Data Description

- Text Review Classification
  - Manual Annotation
  - Automatic Sentence Classification
  - User Reviewing Trends

- Text Based Recommendation System
- Personalized Rating Prediction
- Future Directions
- Conclusions
Annotation Guidelines

• General Guidelines
  ➢ Reviews are annotated at the *sentence level*
  ➢ Each sentence is annotated by one or more topic categories, and only one sentiment
  ➢ No sentence should be considered in the light of previous sentences

• Topic and Sentiment Specific Guidelines
  ➢ Example: For Price Category -
    - Sentences that directly have dollar amounts
    - Sentences that describe portions, value for money and worth.
Manual Annotation

- Three annotators manually annotated 3418 sentences
- 450 sentences for inter-annotator agreement using Kappa Coefficient

<table>
<thead>
<tr>
<th>Sentence Category or Sentiment</th>
<th>K-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>0.88</td>
</tr>
<tr>
<td>Price</td>
<td>0.87</td>
</tr>
<tr>
<td>Staff</td>
<td>0.84</td>
</tr>
<tr>
<td>Ambience</td>
<td>0.62</td>
</tr>
<tr>
<td>Anecdotes</td>
<td>0.51</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>0.60</td>
</tr>
<tr>
<td>Positive</td>
<td>0.83</td>
</tr>
<tr>
<td>Negative</td>
<td>0.78</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.65</td>
</tr>
<tr>
<td>Conflict</td>
<td>0.60</td>
</tr>
</tbody>
</table>
Annotation Example

Example Review
Very romantic fires - I've literally spent hours at Lanterna, drinking wine from their extensive wine and enjoying the ambience. Reasonable prices. HIGHLy RECOMMENDED for a first date. Try the chocolate mud cake (warmed) with 2 scoops of dulce de leche gelato.

Review Sentences
•  <Food><Ambience><Positive><0> Very romantic fires - I've literally spent hours at Lanterna, drinking wine from their extensive wine and enjoying the ambience. </0></Positive></Ambience></Food>
•  <Price><Positive><1> Reasonable prices. </1></Positive></Price>
•  <Miscellaneous><Positive><2> HIGHLy RECOMMENDED for a first date. </2></Positive></Miscellaneous>
•  <Food><Positive><3> Try the chocolate mud cake (warmed) with 2 scoops of dulce de leche gelato. </3></Positive></Food>
Automatic Sentence Classification

• Classification Details
  ➢ Done at the **sentence level**
  ➢ Features are stemmed unigrams

• Trained one Support Vector Machine classifier for each topic and each sentiment

• Classification Evaluation
## Classification Results

<table>
<thead>
<tr>
<th>Sentence Category or Sentiment</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>84.32</td>
<td>81.43</td>
<td>76.72</td>
</tr>
<tr>
<td>Service</td>
<td>91.92</td>
<td>81.00</td>
<td>72.94</td>
</tr>
<tr>
<td>Price</td>
<td>95.52</td>
<td>79.11</td>
<td>73.55</td>
</tr>
<tr>
<td>Ambience</td>
<td>90.99</td>
<td>70.10</td>
<td>54.64</td>
</tr>
<tr>
<td>Anecdotes</td>
<td>87.20</td>
<td>49.15</td>
<td>44.26</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>79.40</td>
<td>61.28</td>
<td>64.20</td>
</tr>
<tr>
<td>Positive</td>
<td>73.32</td>
<td>74.94</td>
<td>76.60</td>
</tr>
<tr>
<td>Negative</td>
<td>79.42</td>
<td>53.23</td>
<td>45.68</td>
</tr>
<tr>
<td>Neutral</td>
<td>80.86</td>
<td>32.34</td>
<td>23.54</td>
</tr>
<tr>
<td>Conflict</td>
<td>92.06</td>
<td>43.96</td>
<td>35.68</td>
</tr>
</tbody>
</table>

**Average classification accuracy for 7-fold cross validation**
Cuisine-Specific Classifier

- Structured metadata has valuable information
- Training and test sets belong to same cuisine

Cuisine-specific classifiers have about 10% improvement over general classifiers

![Average Accuracy across sentence types](chart)

**Average Accuracy across sentence types**

<table>
<thead>
<tr>
<th>Cuisine Types</th>
<th>General Classifier</th>
<th>Cuisine-Specific Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>Chinese</td>
<td>75</td>
<td>85</td>
</tr>
<tr>
<td>Deli</td>
<td>72</td>
<td>82</td>
</tr>
<tr>
<td>French</td>
<td>78</td>
<td>88</td>
</tr>
<tr>
<td>Pizza</td>
<td>80</td>
<td>90</td>
</tr>
</tbody>
</table>
User Reviewing Trends

Review text composition (278245 Sentences)

39% sentences focus on food
67% sentences are positive

Sentence Category

- Food
- Price
- Staff
- Ambience
- Anecdotes
- Miscellaneous

Number of sentences (%)

<table>
<thead>
<tr>
<th>Sentence Category</th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
<th>Conflict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td></td>
<td></td>
<td>39%</td>
<td></td>
</tr>
<tr>
<td>Price</td>
<td>5%</td>
<td>3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staff</td>
<td>15%</td>
<td>10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambience</td>
<td>20%</td>
<td>15%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anecdotes</td>
<td>25%</td>
<td>20%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>30%</td>
<td>25%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As price increases, the number of positive price-related sentences decreases while the number of negative sentences increases.
User Reviewing Trends

Correlation between Text and Star Rating

- Correlation between star rating and the percentage of positive sentences = 0.45
- Star rating and negative sentences = -0.48
- Good reviews mainly have positive sentences (71%), and very few negative sentences (6%).
- Bad reviews have fewer than 5% positive sentences and above 78% negative sentences.
Outline

• Motivations
• Challenges in analyzing text
• Data Description
• Text Review Classification
• Text Based Recommendation System
  ➢ Evaluation Setting
  ➢ Sentiment-Based Text Rating
  ➢ Regression-Based Text Rating
• Personalized Rating Prediction
• Future Directions
• Conclusions
Recommendation Systems

• Users should not have to search reviews
• Users should get recommendations for products that would match some definition of preference
• Current systems rely predominantly on structured metadata, often only star ratings

• How about using the information in the text?
Text-Based Recommendation System: Steps

- Text-derived rating score
  - Sentiment-based rating
  - Sentiment-based rating using topic information
  - Regression-based rating

- Goals
  1. Predicting the metadata star rating
  2. Predicting the text-derived score

- Prediction Strategies
  - Restaurant average prediction
  - Personalized prediction
Sentiment-Based Text Rating

• Translated our annotated text reviews into a **text rating score** using the formula:

  ➢ **Text Rating** = \([P/(P + N) \times 4] + 1\)

    - \(P\) = Number of Positive Sentences
    - \(N\) = Number of Negative Sentences in the review

• The rating is **scaled in the [1,5] interval**.
Text Based Recommendation System: Evaluation Setting

• For evaluation, we separated two non-overlapping test sets of about 260 reviews:
  ➢ Test I: Contained one review for users who have written more than 12 reviews
  ➢ Test II: Users with more than 5 reviews

• For measuring accuracy of prediction we use Mean Square Error (MSE)
Sentiment-Based Text Rating

Restaurant Average-based Prediction

- Prediction using average rating given to a restaurant by all users
- MSE Errors:

<table>
<thead>
<tr>
<th></th>
<th>TEST I</th>
<th>TEST II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicting Star Rating (Goal 1)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Star Rating</td>
<td>1.217</td>
<td>1.295</td>
</tr>
<tr>
<td>Sentiment-based text rating</td>
<td>1.098</td>
<td>1.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TEST I</th>
<th>TEST II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicting Sentiment Text Rating (Goal 2)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Star Rating</td>
<td>1.430</td>
<td>1.342</td>
</tr>
<tr>
<td>Sentiment-based text rating</td>
<td>1.277</td>
<td>1.374</td>
</tr>
</tbody>
</table>

Predicting using text does better than popularly used star rating
Sentiment-Based Text Rating, using Categories

Using Sentence Category

<table>
<thead>
<tr>
<th>Predicting Star Rating (Goal 1)</th>
<th>TEST I</th>
<th>TEST II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment-based text rating</td>
<td>1.098</td>
<td>1.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicting Star Rating</th>
<th>Test I</th>
<th>Test II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>1.215</td>
<td>1.308</td>
</tr>
<tr>
<td>Price</td>
<td>1.377</td>
<td>1.424</td>
</tr>
<tr>
<td>Service</td>
<td>1.531</td>
<td>1.623</td>
</tr>
<tr>
<td>Ambience</td>
<td>1.427</td>
<td>1.559</td>
</tr>
<tr>
<td>Anecdotes</td>
<td>1.57</td>
<td>1.676</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>1.221</td>
<td>1.436</td>
</tr>
<tr>
<td>All but Anecdotes</td>
<td>1.096</td>
<td>1.254</td>
</tr>
</tbody>
</table>

Text Rating = \[
\frac{P}{(P + N)} \times 4 \] + 1

Food and Misc. appear to be most important
Regression-based Text Rating

• Textual rating should encompass all topics with varying importance

**Method**

• We use multivariate linear regression
• Each normalized sentence type [(category, sentiment)-pair] is a variable in the regression
• Dependent variable is metadata star-rating

• Used training set to learn the weights for each sentence type; weights are used in computing text-based score
Regression-based Text Rating

Method

• Using the regression weights can cause scores (raw scores) to lay outside the [1:5] range
• We scale regression-based text ratings to have the same mean and standard deviation values as the star ratings (scaled scores)
• Ignored Neutral and Conflict sentiment classes
Regression-based Text Rating

• Regression Weights

<table>
<thead>
<tr>
<th>Regression Weights</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>4.86</td>
<td>1.53</td>
</tr>
<tr>
<td>Price</td>
<td>1.67</td>
<td>1.59</td>
</tr>
<tr>
<td>Service</td>
<td>2.61</td>
<td>0.51</td>
</tr>
<tr>
<td>Ambience</td>
<td>2.35</td>
<td>2.43</td>
</tr>
<tr>
<td>Anecdotes</td>
<td>3.65</td>
<td>2.02</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>5.17</td>
<td>2.27</td>
</tr>
</tbody>
</table>

Food and Misc. appear to be most important
## Regression-based Text

<table>
<thead>
<tr>
<th>Predicting Star Rating</th>
<th>TEST I</th>
<th>TEST II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Rating</td>
<td>1.217</td>
<td>1.295</td>
</tr>
<tr>
<td>Sentiment-based text rating</td>
<td>1.098</td>
<td>1.27</td>
</tr>
</tbody>
</table>

### Two-Sentiment Regression

<table>
<thead>
<tr>
<th>Predicting Star Ratings</th>
<th>TEST I</th>
<th>TEST II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Rating</td>
<td>1.217</td>
<td>1.295</td>
</tr>
<tr>
<td>Regression-based text rating (scaled)</td>
<td>1.089</td>
<td>1.231</td>
</tr>
</tbody>
</table>
Outline

- Motivations
- Challenges in analyzing text
- Data Description
- Text Review Classification
- Text Based Recommendation System
  - **Personalized Rating Prediction**
- Future Directions
- Conclusions
Personalized Recommendations

Method

• K-Nearest Neighbor (KNN) algorithm for finding similar users
• Pearson Distance function with threshold (threshold = 5) for penalizing users with fewer overlapping restaurants
• For prediction uses average of the K neighbors’ score for the target restaurant
• If a neighbor has not reviewed the target restaurant, resorts to restaurant average-case prediction
### Personalized Recommendations

#### MSE Values

<table>
<thead>
<tr>
<th>Predicting Star Ratings</th>
<th>Test I</th>
<th>Test II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Rating</td>
<td>1.210</td>
<td>1.292</td>
</tr>
<tr>
<td>Sentiment-based text rating</td>
<td>1.098</td>
<td>1.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicting Star Ratings</th>
<th>K</th>
<th>Test I</th>
<th>Test II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Rating</td>
<td>1</td>
<td>1.210</td>
<td>1.292</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.200</td>
<td>1.291</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.194</td>
<td>1.291</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.196</td>
<td>1.291</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.189</td>
<td>1.291</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1.194</td>
<td>1.291</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.200</td>
<td>1.292</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>1.200</td>
<td>1.292</td>
</tr>
<tr>
<td>Two-sentiment regression-based text rating (scaled)</td>
<td>1</td>
<td>1.062</td>
<td>1.268</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.060</td>
<td>1.235</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.071</td>
<td>1.231</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.065</td>
<td>1.230</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.066</td>
<td>1.227</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1.068</td>
<td>1.229</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.075</td>
<td>1.229</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>1.076</td>
<td>1.230</td>
</tr>
</tbody>
</table>
Personalized Recommendations

Percentage Improvements over Baseline

MSE percentage improvement of 2.67% (resp. 0.3%) for Test set I (resp. Test set II) equals RMSE improvement of **1.34%** (resp. 0.15%)
Outline

• Motivations
• Challenges in analyzing text
• Data Description
• Text Review Classification
• Text Based Recommendation System
• Personalized Rating Prediction
• Future Directions
• Conclusions
Future Directions

• **URSA** project tasks (Revisited)
  - Structure Identification and Analysis
  - Text and Structure Search
  - Similarity Search in Social Networks

• Sentence classification improvements
• Text-based search
• Clustering and soft-clustering the sparse data
• Evaluation independent of star rating
Conclusions

- Introduction to URSA project.
- Identify and use topic and sentiment information.
- Study the impact of text-derived information in a recommendation system.
- Our classification approach could easily be translated to other types of datasets…

The URSA project won a Google Research Award in 2008
Thank You!