Topic Modeling and Rating Prediction of Yelp Restaurant Reviews

Jingyi Tan 3032130115 jingyi_tan@berkeley.edu
Cheng Li 3032132052 chengli@berkeley.edu
Zhenqi Wang 3032128646 zhenqi_wang@berkeley.edu

May 5, 2017

Abstract

The goal of this project is to identify hidden topics and to estimate individual rating of each topic in Yelp reviews. We identified 14 topics and their high frequency words by using Latent Dirichlet Allocation (LDA). After using extracted topics to describe an example restaurant and its reviews, we found the subtopics highly relevant to and representative of the original reviews, and provided condensed information of aspects of the restaurant. Then, we fitted a linear regression model by using subtopics to reveal hidden ratings for subtopics. Moreover, we also made prediction about the restaurant's overall rating as well as each customer's rating by using subtopics extracted by LDA as new features. Results have shown that the predictions generated similar (or better) validation MSE compared with commonly used supervised learning algorithms.

Keywords — Latent Dirichlet Allocation, subtopic modeling, rating prediction

1 Introduction

Customers always provide feedback to a business, a product, or a service on websites like Yelp, while potential customers would refer to those reviews to make decision of whether to go to a business or not. For any business on Yelp, we can see all the reviews customers provided, the overall ratings (from one star to five stars) it received, the average price level, etc. Sometimes, it is too time-consuming to go through each review about a business, so people only pay attention to the ratings. However, an overall rating cannot convey the information that led a reviewer to that experience, and people have different standards of rating, so it may be misleading. For example, if a person cares a lot about the restaurants ambience and goes to a restaurant with below-average ambience but high rating because of the taste of its food, this person may feel disappointed. Therefore, a comprehensive rating system is needed to improve customer experience.

In this project, we aim at (1) identifying subtopics involved for each single review, (2) estimating the hidden ratings subtopics for restaurants, and (3) determining the association between the hidden ratings of subtopics and the overall rating given by each customer as well as the overall rating of the restaurant.

2 Dataset

We used the dataset Yelp through its Round 9 of The Yelp Dataset Challenge, which contains 4.1 millions user reviews, 144 thousands businesses and 1 million user information. As we believed that customers and businesses from different cities have different characteristics, we narrowed our analysis down to the reviews of businesses from Charlotte of North Carolina. Among 10,177 businesses in

Charlotte, we only kept those having more than 100 customer reviews, which made our dataset have 80,733 reviews data from 407 restaurants. Following is the structure of a review from the review dataset:

yelp_academic_dataset_review.json

```
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```

For topic modeling, we used the text part of the review dataset. We cleaned texts by decapitalizing and removing stopping words, punctuation and symbols, and all non-noun words. For rating prediction, we only focused on reviews of one restaurant who has the most reviews in Charlotte. In total, we have 1,123 reviews.

3 Methodology

3.1 Tf-idf

Tf-idf, which denotes term frequency-inverse document frequency, is a very useful technique to detect important words in a corpus (a collection of documents). Term frequency (TF) meansures how often the word w_i in document d_j, while the inverse document frequency (IDF) measures how much information the word provides, which is equivalent to the frequency of the word w_i in the corpus. The tf-idf value for a word w_i in document d_j is positively affected by term frequency and negatively affected by document frequency, which can be expressed using the following formula:

$$\text{TF-IDF}(w_i, d_j) = \text{TF}(w_i, d_j) \times \text{IDF}(w_i)$$

The IDF can be smooth by using the formula

$$\text{IDF}_{\text{smooth}}(w_i) = log(\frac{N}{1+n_i})$$

where N is the number of document in the corpus and n_i is the frequency of w_i in the corpus.

In this project, TF-IDF is used in supervised method and tried in unsupervised method. In supervised method, we constructed the TF-IDF matrix and used χ^2 independent test to select top 500 keywords from training set (and their TF-IDF values) to build and tune our models. In unsupervised method (Latent Dirichlet Allocation), TF-IDF matrix is the input for the LDA model.

3.2 Latent Dirichelet Allocation (LDA)

To discover latent topics in each review, we used Latent Dirichlet Allocation (LDA), a topic model that generates topics based on word frequency from a set of documents. LDA assumes that (1) documents contain multiple latent topics; (2) each document is assumed to be generated by a generative process defined by probabilistic model; and (3) each topic is characterized by a distribution over a fixed vocabulary. More specifically, the joint distribution of the hiddens (topics) and observed variables (words) is:

$$p(\phi_{1:K}, \theta_{1:D}, z_{1:d}, w_{1:D}) = \prod_{i=1}^{K} p(\phi_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_d, n|\phi_d) p(w_{d,n}|\phi_{1:K}, z_{d,n})$$

where

$$
\begin{aligned}
\phi_{1:K} &: \text{the topics, each } \phi_k \text{ is a distribution over the vocabulary} \\
&\quad \phi_k \sim \text{Dirichlet}_v(\beta) \\
\theta_{1:D} &: \text{the topic proportion for document } 1:D \\
&\quad \theta_d \sim \text{Dirichlet}_K(\alpha) \\
z_{1:D} &: \text{the topic assignments for document } 1:D \\
&\quad z_d \sim \text{Multinomial}_K(\theta_d) \\
w_{1:D} &: \text{the observed words for document } d \\
&\quad w_d \sim \text{Multinomial}_V(\phi_z)
\end{aligned}
$$

LDA learns the distributions (e.g. the distribution of a set of topics, their associated word probabilities, the topic of each word, and the particular topic mixture of each document) by using Bayesian inference. After repeating the updating process for a large number of times, the model will reach a steady state and can be used to estimate the hidden topics, topic mixtures of each document and the words associated with each topic.

3.3 Prediction Mechanism

We make several assumptions about the ratings of hidden topics and the mechanism of general customers:

(1) Each subtopic found related to a restaurant has an individual hidden rating, and the hidden ratings affect the overall ratings. E.g. for a restaurant with an overall rating 4.0, it might be a result of a 3.0 rated service and 5.0 rated food. And though for some particular customers the individual ratings might not hold, we assume customers share the individual hidden ratings statically.

(2) Each customer's rating for a restaurant is the combination of the hidden ratings of major subtopics found in the customer's review. E.g., when LDA found a customer commenting about the food, service, bar and location of a restaurant with probability 0.35, 0.3, 0.3, 0.05, and the customer gave a rating 4.0, we believe the major factors influencing the rating decision of the customer are food, service, bar. And we made the further assumption each factor has equal weight in the first place, with the overall rating: $4.0 = \frac{1}{3}$ {hidden rating of food} $+\frac{1}{3}$ {hidden rating of service} $+\frac{1}{3}$ { hidden rating of bar }. We used equal weights in our prediction model. However, if performance is questionable we may adjust the weight assignments.

With the assumptions made, we can using linear regression to fit the training data. Major subtopics found in the reviews with probability greater than 0.1 were selected as major influence. Given the original customer ratings and reviews, we are able to estimate hidden ratings of popular topics of the restaurants. Furthermore, we used LDA extracted topics and their predicted hidden ratings to predict the overall rating each customer gave.

We applied MSE as a metric of accuracy as the original ratings are integers and predicted ratings are float. To evaluate the prediction results, we compared the performance with the results from major supervised learning methods – random forest, ordinal logistic regression, and multilayer perceptron with logistic activation function. To perform supervised learning, we used the text part of the review dataset to predict the overall ratings the restaurant received.

4 Results

4.1 Topic Modeling

We used the "nltk" and "gensim" libraries available with Python for texts precessing and training LDA model. When training the model, we only considered number of topics T in a range of $[5, 100]$ and tried multiple other language processing techniques. For example, we used tf-idf to weight training corpus and word stems to replace words with similar meanings, and included more word types like verb. In the end, 15 topics with noun-only corpus without any other reweighting gave the most meaningful and interpretable topics.

4.1.1 Hidden topics discovered by LDA

In the output of 15 topics, we found 14 of them are highly interpretable, and one is ambiguous. (As it will show in the following analysis, the ambiguous one is not preventing us using the topics found by, LDA model). Below are the 14 topics with eight most frequently used words in each topic, ordering from top to the bottom by their frequencies. The name of the topics were made by associating and categorizing the high frequency words of each topic.

14 Topics and High Frequency Words of Each Topic

Breakfast	**American**	**Dinner**	**Dessert**	**Night/Bar 1**	**Night/Bar 2**	**Meat**
Breakfast	Lunch	Dinner	Cream	Beer	Bar	Pork
Coffee	Burger	Night	Dessert	Selection	Place	Chicken
Brunch	Place	Restaurant	Ice	Place	Food	Sauce
Bacon	Chicken	Meal	Chocolate	Night	Area	Flavor
Crepe	Time	Menu	Butter	Food	Time	BBQ
Egg	Salad	Steak	Fruit	Bar	Night	Meat
Morning	Order	Service	Pie	Market	Lot	Mac
Toast	Side	Time	Cake	Wine	Patio	Beef
Service 1	**Service 2**	**Location**	**Japanese**	**Italian**	**Mexican**	**SE Asian**
Food	Food	Location	Sushi	Pizza	Tacos	Thai
Time	Place	Biscuit	Rice	Crust	Chicken	Pho
Service	Service	Center	Roll	Mushroom	Burrito	Chicken
Order	Restaurant	Honey	Spicy	Fox	Salsa	Curry
Manager	Charlotte	Shopping	Buffet	Banana	Cornbread	Tea
Experience	Staff	Shop	Tuna	Store	Buffalo	Yum
Server	Time	Boom	Sauce	Pie	Taco	Rotisserie
Waitress	Quality	Road	Shrimp	Mozzarella	Corn	Breast

Table 1: Topics and High Frequency words

4.1.2 Interpretation of LDA result

To illustrate how LDA conveniently condenses a large number of reviews to a limited number of topics and to evaluate the performance of LDA, we selected a sample restaurant, which has most reviews in the training set, to apply our trained LDA model on. We used the trained LDA model on the

1,123 reviews of the restaurants, extracting subtopics and their probabilities contained in each review. Subtopics with probability lower than 0.1 were filtered out and we have found subtopics counts in the 1,123 reviews ordering by popularity in Table 2.

Topic Frequencies of All Reviews

	Service	Japanese	Night/Bar	American	Dinner	Meat	Dessert	Family	All Others
Count	1007	684	508	488	346	131	81	80	150

Table 2: Topic frequency

From the popular subtopics in the tables, we can guess the restaurant provides Japanese food and American food, has meat dishes and a bar to serve dining/drinking at night, and should be a place for dinner with service.

In the business dataset we found the information of this restaurant as in Table 3:

Example Restaurant

Name	Stars	Categories	Address	City	Review count
The Cowfish Sushi Burger Bar	4	Asian Fusion, Sushi Bars, Burgers	4310 Sharon Rd, Ste X05	Charlotte	1123

Table 3: Example Restaurant

The fact above agrees the description by popular subtopics found by LDA. Furthermore, after using human labor to read several reviews provided in Table 4, we confirmed that the LDA description is very accurate.

4.1.3 "Accuracy" of hidden subtopics found by LDA

As LDA is unsupervised method, we decided to use the trained model on several reviews and compare them manually. We applied the trained LDA on the first 5 reviews of the example restaurant in previous part. Table 4 provides the original reviews, with predicted topics words highlighted. Moreover, the output topics and probabilities in shown as well.

As shown in the table, there is a strong association between the original reviews and predicted topics, which indicates that predicted topics capture the major original topics well. As we read more reviews and predicted topics, we found that for reviews that have very limited words (20 words or less), it's difficult for LDA (and maybe even human brains) to identify a topic that belong to the 14 subtopics, as the comments are generally made about overall dinning experiences. However, for other reviews that are more specific, LDA topics provide accurate description of the original meaning.

First 5 Reviews of the Example Restaurant with Predicted Topics and Probabilities

No	Reviews	Rating	Topics Predicted & Probability
1	Came in on a **Friday evening around 6:30** and was told it would be about a 45min wait, which that was expected. We were seated in less than 30 and had an awesome server named, Chris. Ordered the **crab and lobster spring rolls** which were great! I got the dynamic duo **burger** and my girl friend got the **burgooshi box** which was also great!! I'll be back soon!!	5	American, 0.28 Dinner, 0.24 Night/Bar 2, 0.21 Japanese, 0.12
2	I can't believe it took me two years living in Charlotte to finally eat here! Bonus-We had a great parking space and didn't have to park far at all!! Came here for **my husband's birthday**. We were quoted 25-30 minutes but didn't wait long at all. Maybe 15 minutes. They have a good mix of specialty drinks. I wanted to get a milkshake but was too full to get it. We got the **crab dip**. I saw it was highly recommended on YELP and was very pleased. It's a very large portion. We could barely finish it. The dip was creamy with large chunks of crab in it. **Chips** were large and crispy. That sauce that's on it really made it delicious. Everything mixed together was just amazing!! I got the **shrimp tempura roll**. I was surprised at how big it was. It was awesome, especially with the siracha sauce. I also got the Taste Explosion roll. It was my first roll like that. It was flat with ground beef, jalepenos, onions, some type of amazing sauce with a tomato and cilantro sauce on it with cheese. It was wrapped with seaweed and rice. It was delicious!! It was large and very filling. **Ground beef** was amazing. My husband got the firecracker roll. He said it was one of the best sushi rolls he's ever had. He also said he can tell the ingredients were really good quality and very fresh. I secretly told our server, Bryan that it was my husband's birthday and he made it so special. They came out and sang to him and gave us a **birthday cake**! The cake was huge! Chocolate layers with an awesome layer up top. Bryan was such a great server. He was very helpful and was super attentive. Highly recommended!!	5	Meat, 0.30 Japanese, 0.26 Dinner, 0.24 Family, 0.07 Dessert, 0.05
3	The concept of this place is pretty brilliant! Burger bar, sushi bar, **full bar**? Don't mind if I do. As someone mentioned the menu is kind of outrageously large and that usually means lesser quality and a slower kitchen. I appreciate a restaurant that does a small menu phenomenally. It took about 15 minutes (**3 stops by the waitress**) and finally her recommendation before I placed my order. I ordered the original cowfish bento box with a Philly roll, slider, sweet potato fries, edamame, and cucumber salad. The sweet potato fries were delicious and the cucumber salad is literally (the original definition, not the new ridiculous one) the best I've had at any sushi bar. Everything else was pretty basic and **the service was pretty slow**. They messed up my sister's order. My friend and I were nearly done eating by the time she got her meal. The manager brought it out and comp'ed it. We ordered dessert too! **A tower of lemon pound cake, ice cream, whipped cream, and berries**. It was amazing. The three of us couldn't finish it! After dessert we waited for our check. We were chatting, but after 30 minutes or more we had to ASK for our check. The atmosphere was nice and maybe I'll try going at a less busy time before I make a final ruling on the service. I'll be back.	3	Dessert, 0.17 Service 1, 0.17 Service 2, 0.15 American, 0.15 Night/Bar 2,0.14 Japanese, 0.09
4	We had a great lunch at Cowfish on our visit to Charlotte. Although we had to wait an hour, there was enough in the area to keep us occupied and **the hostesses** were willing to text us rather than carry around one of those buzzers. Once we were seated **the server was quick and friendly**. The menu is huge, but we quickly decided on a few items. My boyfriend got the seared tuna BLT and I had a mix box with a **sushi roll** and a **burger**. We also ordered the fried pickle appetizer -- so good! We'll definitely be back next time we're in town.	4	Japanese, 0.37 Service 2, 0.28 American, 0.22 Service 1, 0.09
5	This is a nice addition to the South Park area. Food was great and the **service was good**. It's a great place for groups and foodies. You can get a great **burger** or some delicious **sushi** all in one. I went in a **Friday night**. The place was packed and at 530 there was already an hour wait. The place seemed clean and the atmosphere was warm. Probably will return.	4	Service 2, 0.45 Night/Bar 1, 0.33 American, 0.09 Japanese, 0.08

Table 4: First 5 sample reviews for sample restaurant

4.2 Subtopics Rating Estimation and Overall Rating Prediction

We used the high frequency topics of the example restaurant as our targets for hidden rating prediction: "Family", "Night/Bar", "Dessert", "Japanese", "Meat", "Dinner", "Service" and "American". We merged topics "Service 1" and "Service 2", "Bar 1" and "Bar 2" together due to similar meanings. When training the linear regression model, the design matrix is a dummy variable matrix indicating which subtopics were significantly involved in the review, and then rescaled into weights added up to 1 for each review. As we do not assume intercept term for the linear regression model and each was given equal weight, so the coefficients in the regression model represent the estimated ratings for hidden subtopics.

4.2.1 Ratings for hidden subtopics

We provided the estimated ratings as follows:

Ratings for High Frequency Topics

	Dessert	Meat	Japanese	Night/Bar	American	Dinner	Service	Family
Rating	4.8	4.6	4.5	4.2	4.3	4.3	4.1	3.3

Table 5: Ratings for high frequency topics

This rating implies our example restaurant provides good food like "dessert", "meat" dishes and "Japanese" food; it might not be the best place for "bar", "dinner" with "service" in town, it's reasonable as this restaurant does not specialize in either bar or diner. Moreover, the low rating of "family" suggests that this restaurant is not a good place for family dinner, which is also reasonable as it comes with a full-bar. As we never have the true values of the ratings of hidden topics, we are evaluating the "accuracy" of the predicted ratings when predicting the overall ratings.

4.2.2 Predicting customer rating and restaurant overall rating

Restaurant Overall Rating
As we've made the assumptions that each subtopic has equal weight affecting the overall rating of a restaurant, the averaged estimated rating of the hidden topics: 4.26 is our estimated overall rating of the restaurant. The true rating of this restaurant shown on Yelp is 4, which has to be an integer, considering the rounding error, we believe there's no reason to reject the assumptions made about rating mechanism and the rating prediction by hidden subtopics ratings works well.

Customer Rating
Use the same assumption, we re-trained an OLS model using the estimated ratings of hidden subtopics with probability over 0.1 of 900 reviews as training dataset to predict the overall ratings. The remaining 223 reviews were treated as validation dataset, and the validation MSE is 0.75 against true values as integers in 1-5. We believe the assumptions are true and the model works well when predicting customer review.

Comparisons
In order to further examine the results of using hidden ratings of LDA subtopics to explain Yelp rating mechanism, we applied random forest, ordinal probit regression with L2 regularization, and multilayer perceptron on the original texts of review data to predict the overall rating. For each algorithm, we picked the 500 most frequently used words from tf-idf and performed 10-fold cross-validation to get the validation MSE. The best validation MSE achieved by each supervised learning algorithm and by the OLS are as follows:

Method	**MSE**
linear regression using rating of subtopics as predictor	0.750
random forest	0.744
ordinal logistic regression (L2 regularization)	0.792
multilayer perceptron (logistic activation function)	0.798

Table 6: Evaluations

As it is shown above, among supervised tf-idf + bag of words methods, random forest using gives us the smallest validation MSE 0.744. And, Ordinal probit regression and multilayer perceptron model have MSEs a little higher than the random forest one. In comparison, OLS using estimated ratings of subtopics by LDA with an MSE at the same level with the supervised methods, we believe the

subtopics found by LDA and their estimated ratings are reasonable. Also, the customer ratings could be explained by ratings of subtopics.

5 Discussions and Conclusions

In this project, we examined the performance of Latent Dirichlet Allocation for generating subtopics in Yelp review. Fourteen hidden topics and their high frequency words are identified in the Results section in Table 1. As discussed in results section, the subtopics given by LDA accurately represent the original meaning of the review. Moreover, subtopics are used to reveal ratings for hidden subtopics. Originally, we want to test the hypothesis that the overall rating can be estimated by the average ratings for each subtopics. Under this assumptions, we used equal weighted OLS to construct a linear regression model to predict overall rating. As we can see from Table 6, the linear regression model successfully predicted the overall rating with MSE of 0.750. Comparing with other three commonly used supervised methods, linear regression with subtopics as features gives a competitive outcome.

However, there are some limitations for our work. Firstly, the choice of filtering threshold probability of 0.1 is not carefully selected. Secondly, assigning equal weights for each significant subtopics in each review might not be appropriate. Thirdly, the LDA method is unable to model topic correlation. These limitations can be resolved by
1. tuning the filtering threshold probability,
2. using the probabilities generated by LDA as weights in our linear regression model,
3. and trying other models such as Correlated Topic Model (CTM) to improve the overall performance.

References

[1] D. Blei, A. Ng, and M. Jordan. *Latent Dirichlet Allocation.* 1991. Journal of Machine Learning Research, 3:9931022, January 2003.

[2] Ramos, Juan. *Using tfidf to determine word relevance in document queries.* Proceedings of the first instructional conference on machine learning. 2003.