Improving Restaurants

by Extracting Subtopics from Yelp Reviews

James Huang, Stephanie Rogers, Eunkwang Joo

Abstract

In this paper, we describe latent subtopics discovered from Yelp restaurant reviews by running an online Latent Dirichlet Allocation (LDA) algorithm. The goal is to point out demand of customers from a large amount of reviews, with high dimensionality. These topics can provide meaningful insights to restaurants about what customers care about in order to increase their Yelp ratings, which directly affects their revenue. We used the open dataset from the Yelp Dataset Challenge with over 158,000 restaurant reviews. To find latent subtopics from reviews, we adopted Online LDA, a generative probabilistic model for collections of discrete data such as text corpora. We present the breakdown of hidden topics over all reviews, predict stars per hidden topics discovered, and extend our findings to that of temporal information regarding restaurants peak hours. Overall, we have found several interesting insights and a method which could definitely prove useful to restaurant owners.

1 Introduction

YELP ratings clearly have a profound effect on the success of businesses as "an extra half-star rating causes restaurants to sell out 19 percentage points more frequently" (increase from 30% to 49%) [1]. But how can a restaurant point out the demands of its customers from a large amount of reviews? We hope to identify what users care about most when writing their reviews, and ultimately determine what certain restaurants are doing right and wrong in order to receive these ratings.

For problems with high-dimensional data, it becomes difficult to extract prominent or relevant features. However, this data will often have a simpler structure: topics in documents, user preferences, themes in discussions, etc. We can approximate these effects by using lower-dimensional models such as LSI or LDA. By breaking these reviews down into latent subtopics using LDA, we are then able to predict a restaurant's star rating per hidden topic. Ultimately these ratings per hidden topic allow us to pinpoint the reasons for a restaurant's Yelp rating, other than food quality. Some latent subtopics that were extracted from Yelp reviews include service, value, decor, and healthiness. Additionally, temporal topics such as breakfast, lunch and dinner also came up in our findings and proved useful for peak hour observations.

2 Related Work

There are many approaches to factor models for discrete data. Among those that deal with dimensionality reduction techniques, Latent Semantic Indexing (LSI) is among the most basic and well-known[4][6]. LSI is an information retrieval technique which uses singular value decomposition to reduce the data to a latent space representation, allowing for more reliable estimation. LSI faced several issues due to the formulation of the probabilistic model, but Hoffman quickly came up with a generative probabilistic model called Probabilistic Latent Semantic Indexing (PLSI), that models each word in the document as a sample from a mixture model [5]. PLSI potentially has problems with overfitting when dealing with small datasets due to the fact that it estimates the probability distribution of each document on the hidden topics independently [2].

We use the Latent Dirichlet Allocation (LDA) factor model to approach the unsupervised learning of factors and topics for the Yelp restaurant review data. This model treats the probability distribution of each document over topics as a K-parameter hidden random variable rather than a large set of individual parameters (K is the number of hidden topics [2]). Alternatively, Laplacian Probabilistic Latent Semantic Indexing (LapPLSI), is an algorithm which models the document space in a more discriminatory manner using nearest neighbors [3].

Alternatively, there have been successful clustering of terms and text documents using non-negative matrix factorization techniques; such as the factoring of 90,000 terms in e-mails to 50 clusters[9]. Among these, Probabilistic Latent Semantic Analysis[8], forms of K-means clustering, Spectral Clustering [7] all provide similar approaches to factor analysis. This would have been a different approach to a latent class model approach that we took and is potentially interesting future work if we wish to compare results from a latent class model to a matrix factorization approach.

The Yelp dataset that we work on has information on reviews, users, businesses, and business check-ins. We specifically focus on all of the restaurant data with regards to each type of information. Related work on this dataset include: predicting the category (e.g. Italian, Spanish, Thai) of a restaurant given a text document; markov chain review generators that

generate reviews automatically; finding the most positive and negative words of a corpus of reviews. Others has also predicted start ratings of reviews using sentiment analysis and predicted buisness categories using clustering[10].

3 Implementation

A. Dataset

This research is performed with the data from the Yelp Dataset Challenge [10]. This dataset includes business, review, user, and checkin data in the form of separate JSON objects. A business object includes information about the type of business, location, rating, categories, and business name, as well as contains a unique id. A review object has a rating, review text, and is associated with a specific business id and user id. We mainly deal with these two types of JSON data objects. Furthermore, we only examine businesses that are of the "restaurant" category and only reviews associated with restaurant businesses. This results in almost 5,000 restaurants, and over 158,000 corresponding reviews. This dataset, specifically the reviews associated with restaurants, will allow us to extract the latent subtopics and pinpoint areas of interest.

B. Tools

We predominantly used Python scripts. Specifically, we used the Gensim Python Library, which is a topic modeling tool for documents. We used PyGal for data visualization.

C. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2] is a Bayesian generative model for text. It is used as a topic model to discover the underlying topics that are covered by a text document. LDA assumes that a corpus of text documents cover a collection of K topics. Each topic is defined as a multinomial distribution over a word dictionary with $|V|$ words drawn from a Dirichlet $\beta_k \sim Dirichlet(\eta)$.

Each document from this corpus is treated as a bag of words of a certain size, and is assumed to be generated by first picking a topic multinomial distribution for the document $\theta_d \sim Dirichlet(\alpha)$. Then each word is assigned a topic via the distribution θ_d, and then from that topic k, a word is sampled from the distribution β_k. θ_d for each document can be thought of as a percentage breakdown of the topics covered by the document.

The topic distribution of a corpus from the LDA model can be found in numerous ways. With the LDA model [2], Blei et al. also present an Expecatation Maximization algorithm that converges to the most likely parameters (word distributions per topic and topic distributions per word). Hoffman et al. present a variation to this Expectation Maximization algorithm which they describe as an Online Learning algorithm for LDA [5]. This is an online Expectation Maximization approach where the parameter learning uses constant time and memory. To discover our latent topics for restaurant reviews we used this online learning approach where reviews were processed in "batches" and the topic model was updated incrementally after processing each batch. While this was not necessary for the Phoenix Arizona dataset we were working on, this approach can then be applied to larger datasets easily.

We find topic models for our text corpus for a range of topic numbers $K \in [10, 500]$ and for $|V| = 10000$. After stopword removal, only the top $10,000$ occuring words by frequency are considered. We found that $K = 50$ gave very reasonable results for our restaurant review dataset. For small topic numbers vocabulary belonging to seperate topics would become grouped into single topics, and for large topic numbers vocabulary that we might associate with a single topic (such as service) would become seperated into several individual topics.

D. Topic Model Example

As an example, we show the breakdown of word distributions over for several topics on a 50 topic LDA model over a 10,000 word dictionary in Table I. Service, for example, is made up of words such as "service," "asked," and "server," with corresponding numbers 4.3, 3.0, and 2.9 which represent the percent that each word makes up of that subtopic. These make up the word distributions for each topic, with each word falling under at least one category as LDA assumes in the first place.

Lunch	**Healthiness**	**American 1**	**Decor**
8.0% lunch	7.0% menu	7.6% potatoes	2.9% patio
7.5% salad	4.4% options	5.5% rib	2.8% inside
6.6% sandwich	2.9% fresh	4.7% mashed	2.7% seating
4.0% chicken	2.7% vegetarian	3.7% prime	2.2% table
Service	**Location**	**American 2**	**Value**
4.3% service	7.9% phoenix	9.5% fish	7.3% portion
3.5% food	3.1% miss	4.9% chips	5.3% price
3.0% asked	2.7% area	4.3% sliders	3.2% small
2.9% server	1.9% town	3.6% son	1.9% huge

TABLE I: Word Distribution of Topics

Furthermore, we show the breakdown of the following one star review in to the above topics in Figure 1.

> "Bummer, we were psyched to have a new burger place. Don't bother- we waited an hour and a half and found out that our waiter "never turned in our order- Uh, what? We won't be back. The patio is too small and the staff is incompetent. No go!"

.

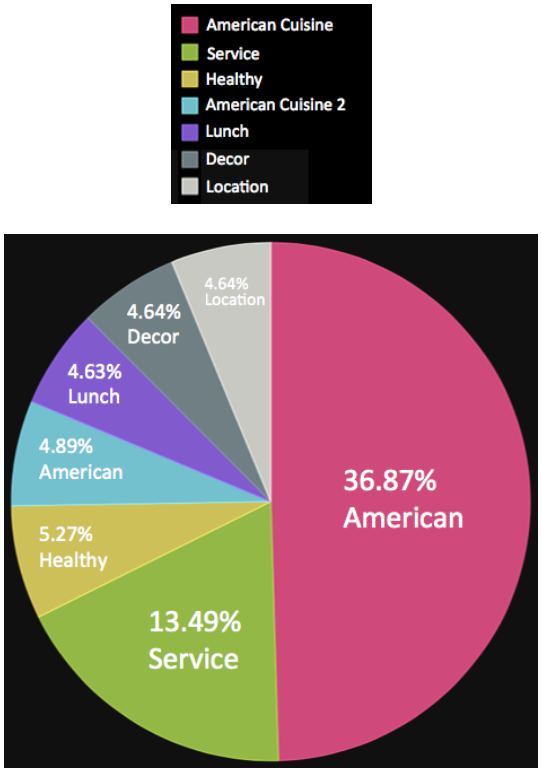

Fig. 1: Example Review Topics

4 Results

A. Hidden Topics

Of the 50 subtopics generated from our Online LDA algorithm, we chose some of the more interesting and more frequently occurring topics to examine. The list of relevant subtopics and the percentage each makes up of all reviews is as follows:

service	8.8%	wait	1.64%
value	5.85%	music	0.77%
take out	3.64%	breakfast	0.59%
décor.	2.99%	dinner	0.50%
healthiness	2.62%	lunch	0.50%

TABLE II: Breakdown of Topics Over All Reviews

According to our algorithm, users care the most about service of all of these subtopics, making up 8.8% of all reviews. Users also care greatly about value, take out and décor. Temporal topics also arise, such as the breakfast, lunch, and dinner categories. These will prove interesting later on when we consider the ratings during these times and compare them to the peak hours of the restaurant.

B. Predicting Hidden Topic Stars

For each restaurant, we predict a star rating per hidden topic above. For example, a restaurant that has an overall rating of 4.0, might have a predicted service star rating of 4.5 and healthiness rating of 3.0. When predicting the star rating per hidden topic, we attempted several different methods. The most basic method we used, was considering all reviews for a restaurant that contained the given topic and averages over all of these review ratings to get the hidden topic rating. We also tried more complex methods including a weighted average using the percentage that the topic made up of the respective reviews, and using the positive and negative weights of neighbor words to those words in the review relating to our given topic. Through manual analysis, we found that the most basic methods proved the most effective.

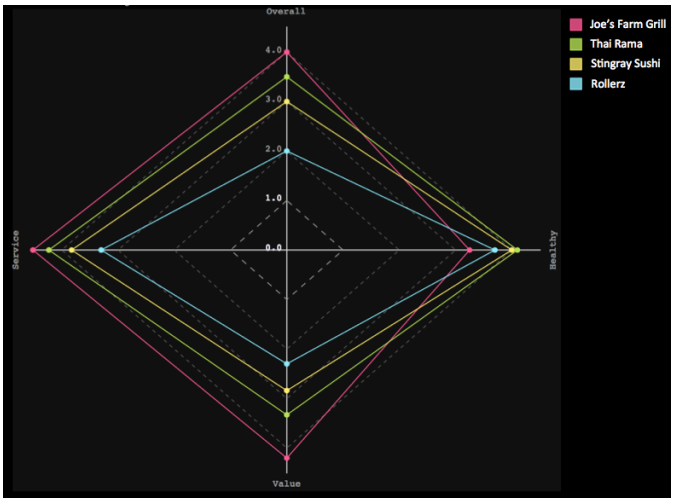

Fig. 2: Four Restaurants' Predicted Subtopic Ratings

Figure 2 is an example of four different restaurants, (Joe's Farm Grill: Burgers, Thai Rama: Thai, Stringray Sushi: Japanese, and Rollerz: Sandwiches) and their respective stars for the service, value and healthiness subtopics.

If we focus on Joe's Farm Grill, we can see that the restaurant has an overall rating of 4.0, a service rating of 4.513, a value rating of 4.203 and a healthiness rating at only 3.25. This means that of the reviews that discussed the healthiness of Joe's Farm Grill, the average rating was lower. In other words, the lower predicted rating from the healthiness subtopic is pulling the overall rating for Joe's Farm Grill down. Based off of this data, we might be able to recommend that Joe's Farm Grill change some of their healthiness choices in order to bring their Yelp rating up.

In order to verify this result, we manually parsed some of the reviews for Joe's Farm Grill which fell under the category of "healthiness." The following 3 star rating review, which contains the healthiness topic, may explain why the healthiness score is tending towards three:

> "The side of veggie fries was literally 3 pounds of fried veggies, full of cholesterol, and way too much for any human to consume."

C. Service Insights

In figure 3, we also show the overall star ratings per restaurant in order to compare it with the predicted subtopic

star ratings. Overall, the average rating of each hidden topic rating is lower if the overall rating of the restaurant is lower. This is explained by the fact that we only average across reviews for the given restaurant, and is accurate because the topics will effect the overall rating in the end. We also found that the quality of food is highly correlated with the quality of service and the quality of many other topics. We performed bigram LDA and received topics that included "great food" and "great service" together as well as "bad food" and "service bad" together. This shows that reviews that talk about the quality of food, tend to also mention the service in an equally positive or negative lighting. This can be explained by halo effect and cognitive bias: if a user thinks the food is good, the service is good by default.

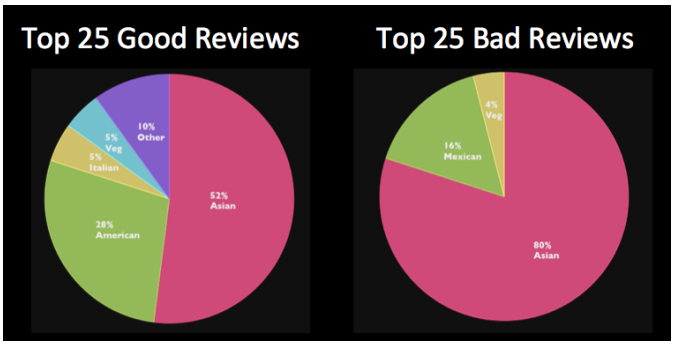

Fig. 3: Reviews in Service

We went through the reviews, and found the top 25 best and worst reviews that dealt with service, according to our service category from our LDA algorithm. The breakdown of types of restaurants that make up the top 25 best and worst reviews is shown in figure 4. We can see that asian food restaurants make up a majority of the top 25 best and worst, with Thai restaurants making up 45% of both the best 25 and worst 25. However, it is interesting to note that there are only 150 Thai restaurants out of the 4,503 restaurants that make up our dataset. This means that Thai restaurants are extremely polarizing in their reviews. Furthermore, we can claim that Western cuisine restaurants care enough to stay off of the worst service list, while making up a majority of the best service reviews.

While manually reading through the 25 worst service reviews, we noticed several mentions of the word "Groupon." Upon further analysis, we have found that there are ten times more mentions of the word "Groupon" in bad service reviews, than there are in good service reviews. This might be explained by the fact that these restaurants are attempting to attract customers and promoting their restaurants through Groupon deals, as they are obviously doing something wrong. However, we can identify their issues as predominantly service, and perhaps even other areas. Groupon simply isn't the right way to be dealing with their issues, they should instead be focusing on raising their Yelp ratings.

D. Temporal Insights

We originally planned to create a recommendation system based on temporal data. We wanted to find interesting restaurants for happy hour, or specifically best breakfast places. As we extracted subtopics which catered to these temporal areas, we decided to compare the average across all reviews with the subtopics of breakfast, lunch, and dinner with the checkin data for breakfast, lunch, and dinner time. With the checkin data, we were able to determine which of those three times were the busiest for the restaurant. We believe the busiest time period probably represents the most popular time period (for example, iHop would be busiest and most popular in the morning, and a sandwich place is most popular and busiest during lunch). However, in comparing the breakfast, lunch, and dinner scores with the checkin data, we found that only 23% of restaurants are rated the highest during peak busy hours, or when it is most popular.

After further investigation, we found that on average, restaurants are actually rated 0.4 lower than their best breakfast, lunch or dinner score than when they are busiest. As the restaurant is busier, the wait time and service may be slightly worse, factors which clearly have an effect on the rating of a review based off of the hidden topics we extracted.

5 Future Work

As we already performed a bigram LDA, it would be interesting to apply the bigram LDA topics in a similar way as the above unigram applications. Furthermore, it would be interesting to use these topics we already found as features in some other algorithm for different purposes.

Since this is unsupervised learning, we are extremely interested in developing a way to determine the accuracy of our predicted hidden topic stars per restaurant. Through the little manual analysis we were able to accomplish, we believe these star are representative and helpful in determining what restaurants are doing right and wrong, and how they can attempt to curb their scores event more. It would be beneficial to have users rate the restaurants based off of some of these topics, and perform some sort of supervised learning classification tool. Additionally, we could also determine which method of predicting stars on an unsupervised set is most accurate.

For our worst and best service reviews, we only identified those within the subset of of 1 and 5 star reviews. We could look into somehow determining the worst reviews with regards to service by weighting the amount of service within the review, to the overall star rating of the review. Ultimately, figuring out a way to include other starred ratings in determining which reviews are best and worst per hidden topic or proving that the 1 star and 5 star reviews are in fact the best and worst would prove useful.

6 Conclusion

Based off of the Online LDA algorithm, we have been able to show what users care about most in their reviews of restaurants, and have been able to pinpoint the areas of interest for specific restaurants. Overall, it turned out that users care most about service, and subsequently value, take out, and decor. Based on the topics we have found, we predicted stars of hidden topics. Those stars varied around the range of the overall rating of the restaurant, as we expected, with lower and higher ratings in certain areas. With these ratings of specific subtopics that Yelp users care about, restaurants could earn insights on how to improve their businesses. Another finding from the review analysis is the change of customer satisfaction based on time. We found that users are less likely to rate high stars during peak times. Through future works, we expect to explore more accurate and specific insights, possibly beneficial to restaurants, from a large amount of reviews.

7 References

[1] M. Anderson and J. Magruder. "Learning from the Crowd." *The Economic Journal*. 5 October,2011.

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

[3] D. Cai, Q. Mei, et all. "Modeling Hidden Topics on Document Manifold." Department of Computer Science, University of Illinois. CIKM 2008.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. "Indexing by latent semantic analysis." *Journal of the American Society of Information Science*, 41(6):391407, 1990.

[5] M. Hoffman and D. Blei. "Online Learning for Latent Dirichlet Allocation." Neural Information Processing Systems, 2010.

[6] C. Papadimitriou, P. Raghavan, et all. "Latent Semantic Indexing: A Probabilistic Analysis." *Journal of Computer and System Sciences*. October 2000.

[7] R. Zass, A. Shashua "A Unifying Approach to Hard and Probabilistic Clustering". International Conference on Computer Vision (ICCV) 2005.

[8] E. Gaussier C. Goutte "Relation between PLSA and NMF and Implications" ACM SIGIR conference on Research and development in information retrieval. SIGIR 2005

[9] B. Murray, . "Email Surveillance Using Non-negative Matrix Factorization". Computational and Mathematical Organization Theory 2005

[10] https://www.yelp.com/academic_dataset