# 1. Background of customer review analysis and its business impact

To be added

# 2. Data Description

The data in this project was taken from Yelp data challenge round 7, which includes 67584 US businesses and 9858 business in other countries. The type of business includes restaurants, shopping centers, super markets, Home Services, etc, while restaurant is the largest group and accounts for around 25% of population. The data for restaurants around university of Illinois, Champaign and Urbane (UIUC) were studied in this research, around 300 business and 11205 reviews.

Besides customer views, Yelp also provides comprehensive business attributes such as open hours, neighborhood, categories.

**yelp\_academic\_dataset\_business.json**

{

"business\_id":"encrypted business id",

"name":"business name",

"neighborhood":"hood name",

"address":"full address",

"city":"city",

"state":"state -- if applicable --",

"postal code":"postal code",

"latitude":latitude,

"longitude":longitude,

"stars":star rating, rounded to half-stars,

"review\_count":number of reviews,

"is\_open":0/1 (closed/open),

"attributes":["an array of strings: each array element is an attribute"],

"categories":["an array of strings of business categories"],

"hours":["an array of strings of business hours"],

"type": "business"

}

**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"

}

# 3. Exploratory Analysis

The geographic distribution of restaurants can be shown by plotting the latitude and longitude on the google map, as shown in Figure 1, where size of dot indicates the rating stars. It could be found that restaurants are located around the UIUC campus.

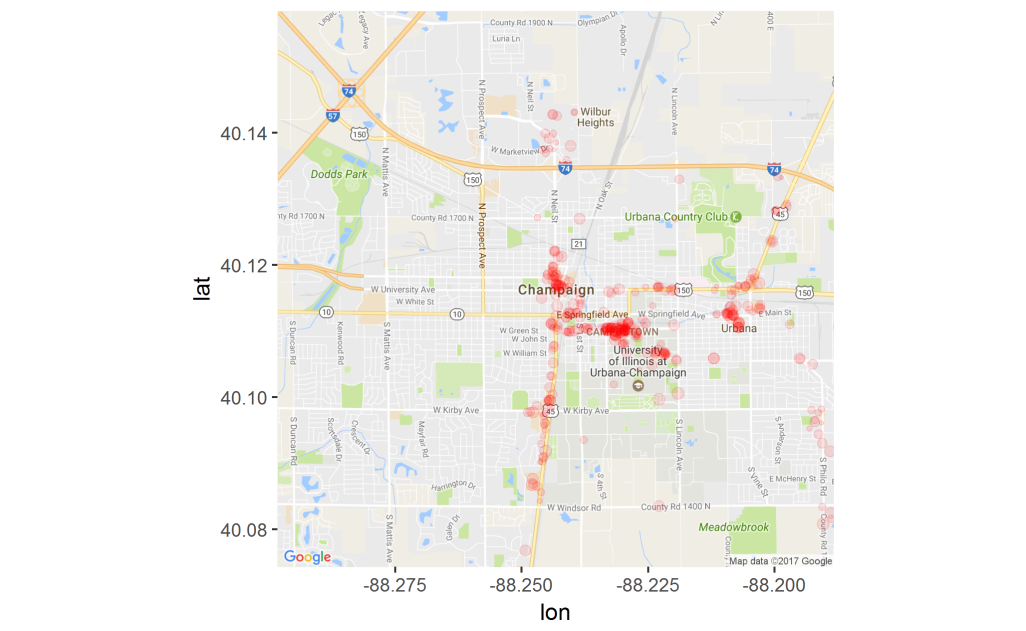


Figure 1 geographic distribution of restaurants around UIUC in Yelp Dataset

The reviews were rated from 1 to 5, while the overall rating for a business is rounded to every 0.5. The rating distribution among reviews and restaurants are shown in Figure 2, which indicates that customers generally leave positive reviews.

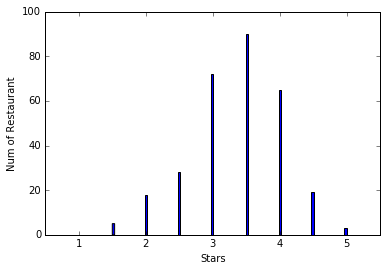
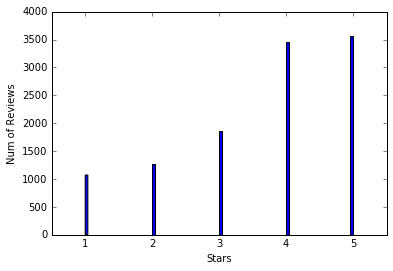


Figure 2 Rating star distributions of restaurants in the studied reviews a) rating of individual reviews b) average review ratings of restaurants

Figure 3 plots the relationship length of review text in words, binned by an interval of 20, against the average rating for reviews of that particular length. The trend that is observed indicates that there is a steady decrease in the average rating as the number of words in the review increases. Moreover, the variation in average rating for lengths that are approximately the same is seen to be extremely high as the length increases, especially beyond 500, when compared to lower to middle lengths.

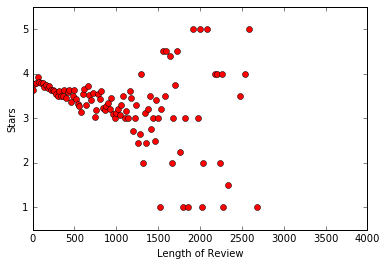
`

Figure 3 relationship between length of review and average rating

# 4. Preprocessing

## 4.1 Tokenizing sentences into bag of words:

In text mining, documents are simply represented as bags of words. Each sentence can be converted into a vector of distinct terms and terms frequency.

## 4.2 Removing stop words and punctuations:

Stop words refer to the most common words in the language which does not provide too much information for the document, such as ‘we’, ‘the’ and ‘there’; Stop words were removed against the NLTK stop words dictionary, which includes 153 words for English language. Common punctuations were also removed during this process.

## 4.3 Lemmatization

Words have different forms, such as ‘do’, ‘did’ and ‘doing’, and different derivations, such as ‘memory’ and ‘memorize’. Lemmatization refers to converting different forms or derivation of terms to the base or dictionary form of a word, which is known as the lemma.

# 5. Topics Modeling and Feature Extraction

Two approaches, bigram/trigram phrase extraction and Latent Dirichlet allocation (LDA) were used for topic modeling.

## 5.1 Bigram/Trigram Topic modeling

In text mining, bigram is a sequence of two adjacent words while trigram is a sequence of three. E.g, ‘machine learning’ and ‘Chinese food’. Extracting bigram or trigram is a relative convenient way to fetch topics within words which is easy to interpret.

Use the GENISM wordtovec function, the frequent bigram/ trigram phrases were extracted from the review as table x.

Table 1 the 50 most common bigram and trigram phrase in studied reviews

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bigram Phrases | Frequency |  | Bigram Phrases | Frequency |
| pretty\_good | 629 |  | pulled\_pork | 240 |
| food\_good | 572 |  | love\_place | 232 |
| good\_food | 361 |  | fried\_rice | 226 |
| great\_place | 333 |  | service\_good | 216 |
| black\_dog | 326 |  | beer\_selection | 215 |
| champaign\_urbana | 308 |  | staff\_friendly | 214 |
| 5\_star | 299 |  | highly\_recommend | 210 |
| chinese\_food | 284 |  | good\_service | 209 |
| mexican\_food | 276 |  | service\_great | 208 |
| food\_great | 272 |  | good\_place | 206 |
| great\_food | 267 |  | fast\_food | 200 |
| burnt\_end | 262 |  | great\_service | 189 |
| deep\_dish | 244 |  | quality\_food | 183 |
| sweet\_potato | 244 |  | chip\_salsa | 180 |
| pad\_thai | 243 |  | customer\_service | 173 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trigram Phrases | Frequency |  | Trigram Phrases | Frequency |
| sweet\_potato\_fry | 162 |  | big\_grove\_tavern | 28 |
| deep\_dish\_pizza | 87 |  | highly\_recommend\_place | 27 |
| food\_pretty\_good | 75 |  | hot\_sour\_soup | 26 |
| champaign\_urbana\_area | 74 |  | thai\_iced\_tea | 25 |
| thin\_crust\_pizza | 69 |  | 3\_5\_star | 25 |
| chicago\_style\_pizza | 52 |  | pad\_kee\_mao | 25 |
| baked\_potato\_casserole | 47 |  | chicken\_pad\_thai | 24 |
| pulled\_pork\_sandwich | 41 |  | 4\_5\_star | 24 |
| authentic\_mexican\_food | 35 |  | beer\_battered\_bacon | 23 |
| general\_tso\_chicken | 34 |  | burnt\_end\_sandwich | 22 |
| bi\_bim\_bap | 34 |  | sweet\_sour\_chicken | 22 |
| place\_champaign\_urbana | 34 |  | sweet\_potato\_chip | 22 |
| great\_beer\_selection | 32 |  | service\_bit\_slow | 21 |
| restaurant\_champaign\_urbana | 30 |  | free\_chip\_salsa | 21 |
| authentic\_chinese\_food | 30 |  | give\_5\_star | 21 |

It can be found that these phrases provided comprehensive information about the content of customer commons. The phrase like ‘food\_good’, ‘great place’,‘give\_5\_star’ actually directly provide the customers’ feedback.

These bigram/trigram phrases could be used as self–generated labels when new reviews are added. For example, if a business has matched phrases, it can be represented by those labels. They also could be used as feature for predictive models.

## 5.2 LDA (Latent Dirichlet allocation) Model

LDA is a generative statistical model for text mining, which considers each document as a mixture of topics. It uses a three-level of hierarchical Bayesian model to fit the data. A fixed number of topics could extract from all the reviews, and then a particular review could be represented as a mixture of topics. Take one review for example:

“I think this one is actually a 3.5 star, if that were allowed...An unpretentious little neighborhood spot, this bar and grill has a 'sports bar' theme but a very relaxed, although almost too clean for, near-dive-y feel. Not crammed with student crowds and a good mix of ages although many boomer types were represented. On U of I basketball nights, they serve free chili. Our dinner plans canceled abruptly, two housemates and I headed over to check it out. Expecting a watery, bean heavy chili soup; I was pleasantly shocked by the meaty and rich actuality. It was sooo good that we all decided to stop by sometime to try menu items we'd actually pay for in cash! Drinks were good, and strong after they realized we'd tip. The bar was crowded, but not unpleasantly. Only complaint was that the fresh diced onion and grated cheese toppers were not refilled after they ran out.”

It could be represented as a mixture for following topics out of 50 extracted topics:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Topic (10 most probable terms out of 19175 dictionary) | Proportion | Manual Interpretation |
| 29 | 0.130\*beer + 0.088\*bar + 0.040\*drink + 0.038\*selection + 0.026\*place + 0.021\*good + 0.021\*bartender + 0.020\*game + 0.018\*night + 0.018\*tap | 13.9% | Bar |
| 11 | 0.033\*italian + 0.014\*walking + 0.013\*chili + 0.011\*turkey + 0.010\*fault + 0.010\*husband + 0.010\*7 + 0.009\*lunch + 0.009\*pasta + 0.009\*opted | 12.1% | Food-Chili |
| 39 | 0.045\*place + 0.038\*nice + 0.024\*table + 0.023\*seating + 0.022\*good + 0.018\*great + 0.018\*staff + 0.017\*area + 0.016\*spot + 0.015\*sit | 13.1% | Service-(Seating) |
| 31 | 0.031\*table + 0.024\*time + 0.024\*waitress + 0.022\*asked + 0.021\*food + 0.019\*server + 0.018\*didn + 0.018\*service + 0.017\*drink + 0.015\*ordered' | 11.3% | Service  (waiting) |
| 45 | 0.066\*cheese + 0.049\*salad + 0.034\*bread + 0.028\*good + 0.024\*sandwich + 0.018\*tomato + 0.016\*onion + 0.015\*soup + 0.015\*chicken + 0.013\*lunch | 9.8% | Food-cheese |

It could be found from this model, as LDA is a generative statistical model and a topic is presented as distribution of 10 most probable terms from a large dictionary, the topics are not as straight forward to interpret as the bigram/trigram phrases. Sometimes, no words from the 10 most probable words actually appear in the real text. However, the advantage of LDA is to provide a way to break down the document into different conceptualized topics.

An important parameter of LDA is the number of topics to be derived. When the number is small, information from different topics could combine into one, while one topic could be divided into multiple subtopics when number is large. A topic number 50 was chosen after trials with 5, 20, 50 and 100.

# 6. Sentimental Analysis

Sentimental polarity of customers was evaluated based on a [-1,1] scale; where -1 represent absolute negative, 0 as neutral and 1 means positive. The polarity score of each sentence was derived using VADER approach. The VADER (Valence Aware Dictionary for Sentiment Reasoning) algorithm provides a polarity score lookup table of sentiment related words, which was derived from training on social media text(tweets, movie review and amazon reviews) . Then the polarity score of each review is derived by averaging of polarity scores of sentences, the distribution of which is shown in Figure 4.

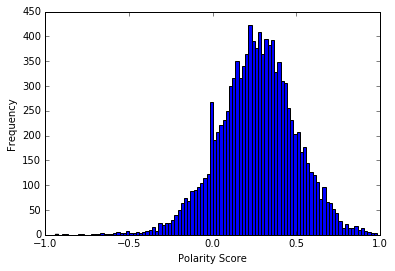


Figure 4 distributions of derived sentiment polarity/intensity scores

# 7. Predictive Analysis

Review ratings in a range of 1 to 5 can be considered either numerical or categorical, as there is no distinct difference when a 1 star review is misclassified as 2, while it is a terrible error if it is labeled as 5. Both approaches were implemented and compared. Regression models were built for numerical ratings; According to methodology of ‘Net Promotor Score’, an industrial system for measuring customer experience, Ratings 1 to 3 were relabeled as ‘detractor’, 4 as ‘indifference’ and 5 as ‘promoter’ for building the classification models.

**7.1 Regression Analysis of Review Ratings**

Linear regression, lasso, support vector machine, and random forest methods are used to build the regression models. 5 folder cross validation were used to evaluate the model performance in term of R squared, as shown in Table 2. Simple linear regression yielded a large negative R2 which and support vector machine, while the random forest achieved highest performance.

Table 2 Comparison of regression model performance based 5 fold cross validation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | 5 fold R2 | | | | | Average R2 |
| Linear Regression | 0.3896 | 0.3083 | 0.3504 | -6.104 | 0.4229 | -0.92656 |
| Lasso | 0.2170 | 0.2299 | 0.2151 | 0.1537 | 0.2430 | 0.21174 |
| SVR | -0.0712 | -0.0027 | 0.0584 | 0.0281 | 0.0321 | 0.0089 |
| Random Forest | 0.3813 | 0.3189 | 0.3761 | 0.3217 | 0.4460 | 0.3688 |

Random Forest model allows evaluating the relative importance of variables in the model building. Figure 5 shows the 50 most important features. It could be found that most important attribute is the derived sentiment polarity score, while the length of review is the second. The rest of import features are the topics derived from LDA approaches. Although LDA derived topics are harder to interpret, obvious they capture more useful information than the bigram and trigram phrases.

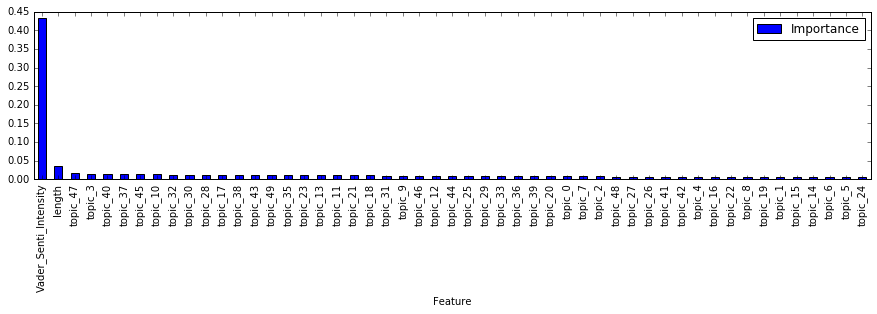


Figure 5 distributions of derived sentiment polarity scores

More model tuning and feature selection should be done in the near future

**7.2 Classification Analysis of Review Ratings**

Logistic regression, support vector machine classification (SVC) , and random forest classification are used to build the classification models to predict the NPS category of rating, ‘Promotor’, ‘Indifference’ and ‘Detractor’. Surprisingly, logistic regression actually achieved the highest accuracy. We could calculate the accuracy of regression model by converting the numerical prediction from into NPS category, which is 0.5331 for random forest regression, which is still lower than the accuracy of Logistic regression.

Table 3 Comparison of classification model performance based 5 fold cross validation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | 5 fold Accuracy | | | | | Average |
| Logistic regression | 0.5660 | 0.5368 | 0.4908 | 0.5443 | 0.5924 | 0.5460 |
| SVC | 0.3898 | 0.3922 | 0.3837 | 0.3908 | 0.4107 | 0.3934 |
| Random Forest | 0.4924 | 0.4859 | 0.5227 | 0.4796 | 0.4859 | 0.4859 |

More model tuning and feature selection should be done in the near future