



Disaggregation of Yelp Review Ratings

Peiqing Lian, Xinrong Lian, Zelong Chen



Motivation

*I'm gonna crack into
all the reviews and
see what's going on
there. That's what a
genius should do*






Rockstar or Five star?

Problem Statement



- Use Categorical sentiment intensity scores as predictor values to predict which category would contribute the most to star rating.

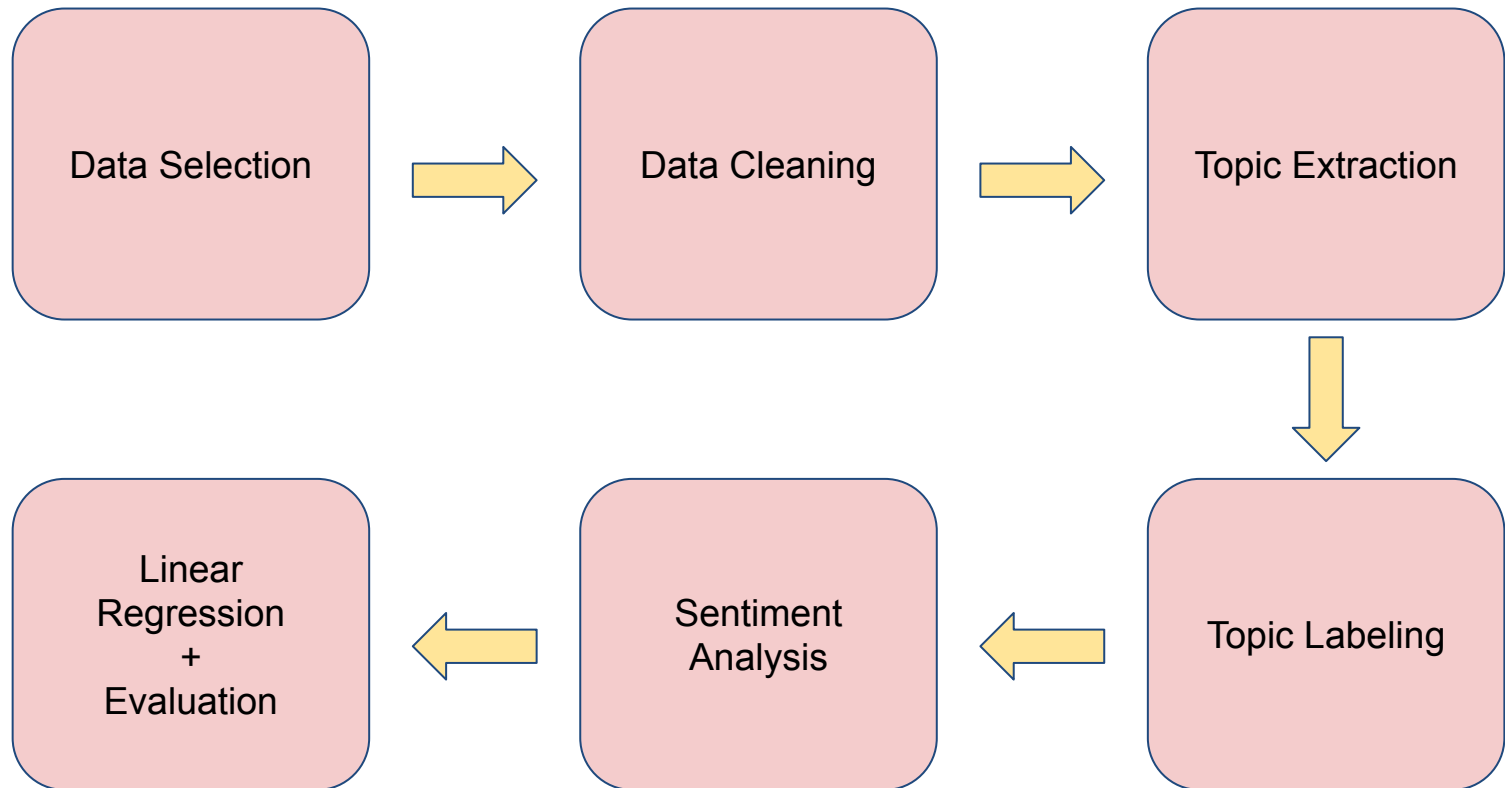
Users	Predictors				Target
	Waittime	Service	Food	...	
	+0.3	+0.4	+1.4	...	★★
	+0.7	+1.2	+0.5	...	★
	+2.1	+0.8	+1.5	...	★★★

Yelp Academic Dataset



- **Yelp Academic Dataset**
 - 6,685,900 Reviews & 192,609 Businesses
 - [Yelp.com/dataset/challenge](https://www.yelp.com/dataset/challenge)
- **Restaurants in Las Vegas, Nevada (1,242,711 Reviews)**
 - **By Postal Code**
 - 89103, 89109, 89118, 89119, 89169
 - 697,325 Reviews
 - **By Restaurant**
 - Hash House A Go Go
 - 5,847 Reviews

Solution (Overview)



Topic Extraction



- **Data Cleaning**
 - **Tokenize Reviews by Sentences**
 - **Remove Sentences with Length Less than 10**
 - **Lowercasing**
 - **Remove Common English Stop Words**
- **TF-IDF Vectorizer**
 - **Input: Corpus**
 - **Documents: Tokenize Sentences from Reviews**
 - **Output: Matrix of the Document Feature Vectors**

Topic Extraction



- **Non-Negative Matrix Factorization (NMF)**
 - **Dimensionality Reduction on the TF-IDF Matrix to Obtain NMF Matrix**
 - **Number of Topics: 7 Number of Top Words: 15**

Topics in NMF model (generalized Kullback-Leibler divergence):

Topic #0: great service friendly excellent experience staff customer slow server fast atmosphere attentive waiter quick bad

Topic #1: chicken waffles fried sage benedict ordered bacon got eggs delicious andy waffle potatoes crispy hash

Topic #2: huge portions large big share portion delicious people prices plate massive enormous hungry meal tasty

Topic #3: good really pretty service overall just potatoes biscuits bloody thing mary taste coffee biscuit wasn't

Topic #4: place vegas breakfast definitely hash love house try time come eat best recommend just last

Topic #5: food amazing delicious man vs awesome just came lot price excellent took quality tasty large

Topic #6: wait worth long time minutes hour seated 30 table minute 45 20 come definitely 10

- Topic #0: Service
- Topic #1: Food
- Topic #2: "Worth it"
- Topic #3: Food / Service
- Topic #4:
- Topic #5: Food
- Topic #6: Wait

Topic Labeling



- **Non-Negative Matrix Factorization (NMF)**
 - Rows in NMF matrix = Tokenize Sentences from Reviews
 - Label Sentences with the Topic with the Max NMF Score

Sample Review:

Massive is an understatement!

Sample Review:

The chicken was crispy and the best part was finding bacon pieces in my waffle.

Topic Distribution:

Topic 2: 0.0032

Topic 5: 0.0016

Topic 1: 0.0007

Topic 4: 0.0002

Topic 0: 0.0000

Topic 3: 0.0000

Topic 6: 0.0000

Topic Distribution:

Topic 1: 0.0523

Topic 4: 0.0040

Topic 3: 0.0006

Topic 0: 0.0000

Topic 2: 0.0000

Topic 5: 0.0000

Topic 6: 0.0000

Sentiment Analysis



- **VADER Sentiment Analysis**
 - **Returns the Sentiment Intensity Score of the Sentence**

I would definitely recommend this place to others and I will be back next time I'm in Vegas!
{ 'neg': 0.0, 'neu': 0.718, 'pos': 0.282, 'compound': 0.6696 }

Waited only about 15 minutes to be seated, though.
{ 'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0 }

This place freaking rocks.
{ 'neg': 0.483, 'neu': 0.517, 'pos': 0.0, 'compound': -0.4215 }

103	10	Bumping to four stars.. the food is just too g...	4	3	0.4404
104	10	Cell service is still non-existent and the wif...	4	0	-0.3724
105	10	Chicken Benedict really is one of the best thi...	4	1	0.6369

userid	2923.000000
0	0.145195
1	0.123677
2	0.165368
3	0.173203
4	0.176851
5	0.190683
6	0.086749
stars	3.919446

- **Aggregate Sentiment Intensity Scores by Averaging**

10	10.0	-0.3724	0.636900	0.000000	0.44040	0.000000	0.000000	0.000000	4.0
----	------	---------	----------	----------	---------	----------	----------	----------	-----

Linear Regression+Evaluation



- Simple Linear Regression:

$$\hat{y} = 3.17 + 0.68 * Service + 0.50 * Food1 + 0.53 * Worth + 0.28 * Food/Service + 1.12 * Topic4 + 0.97 * Food2 + 0.69 * Wait$$

Mean Squared Error: 1.0871000026560464

AIC: 13435.641863966863

- Removing Intercept and Uncleared Topics:

$$\hat{y} = 2.41 * Service + 2.61 * Food1 + 3.11 * Worth + 2.82 * Food/Service + 3.40 * Food2 + 2.47 * Wait$$

Mean Squared Error: 6.117380574771179

AIC: 21352.176868844

- Trade-Off Between Accuracy vs Interpretability

Linear Regression+Evaluation



- **Cross-Validation (10% Dataset)**

- **Best Model:**

- **Mean Square Error:1.034**

```
models      MSE
2 Ridge  1.034118
[{'normalize': True}, {'alpha': 0.1}, {'alpha': 1}, {'alpha': 0.1}]
```

- **Test Results (10% Dataset)**

- **Best Model:**

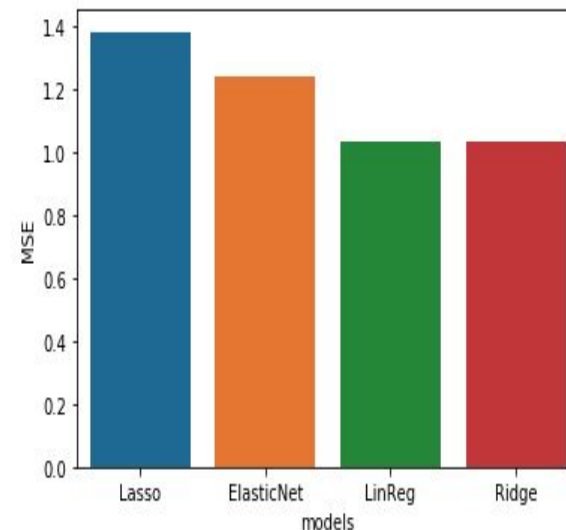
- **Mean Square Error:1.0557**

- **AIC: 12958**

- **Simple Regression:**

- **Mean Square Error:1.0558**

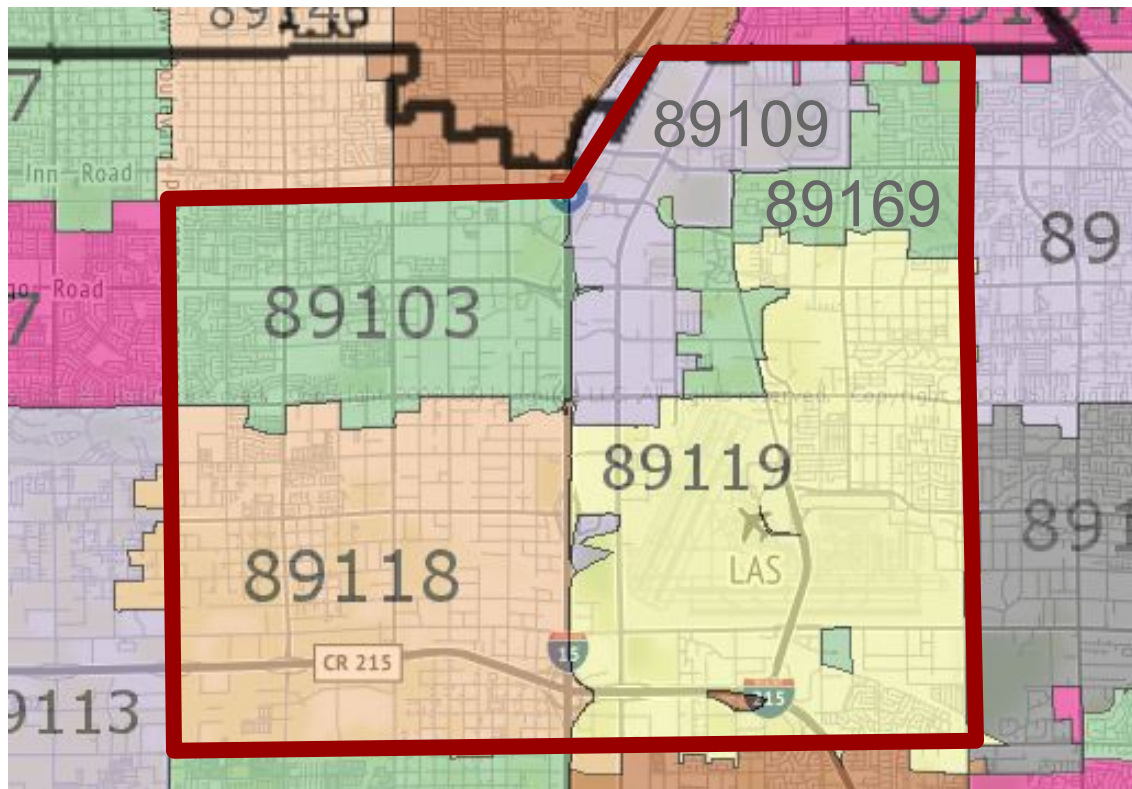
- **AIC: 12957**



Restaurants in Las Vegas



- Visualization By Postal Code



Restaurants in Las Vegas



- Equations by Postal Code

Postal Code 89119

$$\hat{y} = 2.79 + 0.69 * \text{Service} + 0.39 * \text{Food1} + 0.84 * \text{Worth} + 0.95 * \text{Food/Service} + 1.02 * \text{Topic4} + 1.84 * \text{Food2} + 0.61 * \text{Wait}$$

Postal Code 89103

$$\hat{y} = 2.76 + 0.67 * \text{Service} + 0.49 * \text{Food1} + 0.82 * \text{Worth} + 1.06 * \text{Food/Service} + 1.00 * \text{Topic4} + 1.83 * \text{Food2} + 0.67 * \text{Wait}$$

Postal Code 89118

$$\hat{y} = 2.99 + 0.57 * \text{Service} + 0.48 * \text{Food1} + 0.75 * \text{Worth} + 0.88 * \text{Food/Service} + 0.95 * \text{Topic4} + 1.70 * \text{Food2} + 0.71 * \text{Wait}$$

Postal Code 89169

$$\hat{y} = 2.96 + 0.59 * \text{Service} + 0.36 * \text{Food1} + 0.68 * \text{Worth} + 0.86 * \text{Food/Service} + 0.94 * \text{Topic4} + 1.66 * \text{Food2} + 0.77 * \text{Wait}$$

Postal Code 89109

$$\hat{y} = 2.87 + 0.63 * \text{Service} + 0.32 * \text{Food1} + 0.85 * \text{Worth} + 0.84 * \text{Food/Service} + 0.90 * \text{Topic4} + 0.75 * \text{Food2} + 1.65 * \text{Wait}$$



Future Work

- Latent Dirichlet Allocation (LDA)
 - Dimensionality Reduction
 - Alternative to Non-Negative Matrix Factorization (NMF)
- Alternative to VADER Sentiment Analysis

```
Waited only about 15 minutes to be seated, though.  
{ 'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0 }
```

```
This place freaking rocks.  
{ 'neg': 0.483, 'neu': 0.517, 'pos': 0.0, 'compound': -0.4215 }
```