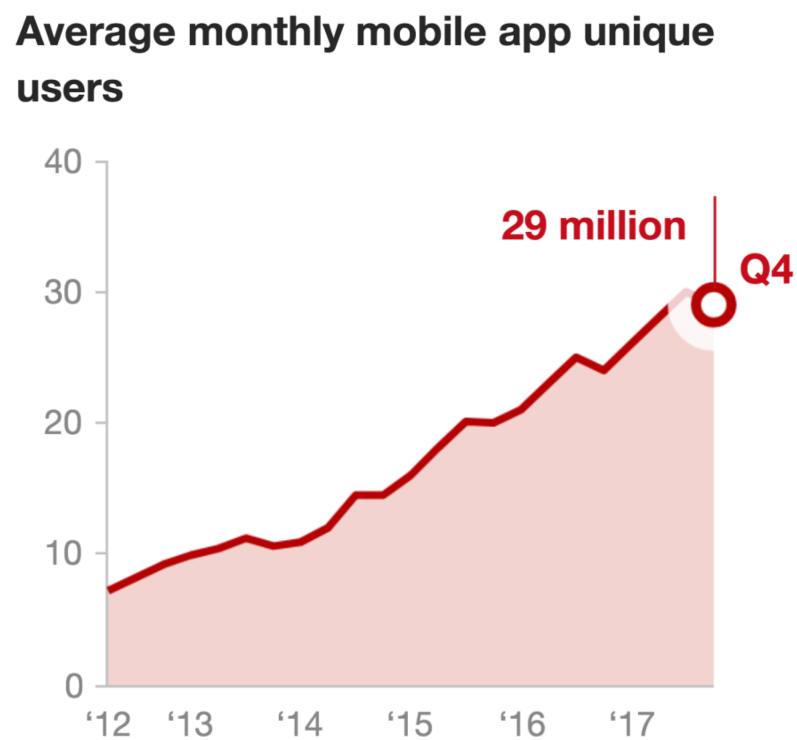
Should I invest it?
Predicting future success of restaurants using yelp dataset

Xiaopeng Lu, Jiaming Qu

PEARC' 18

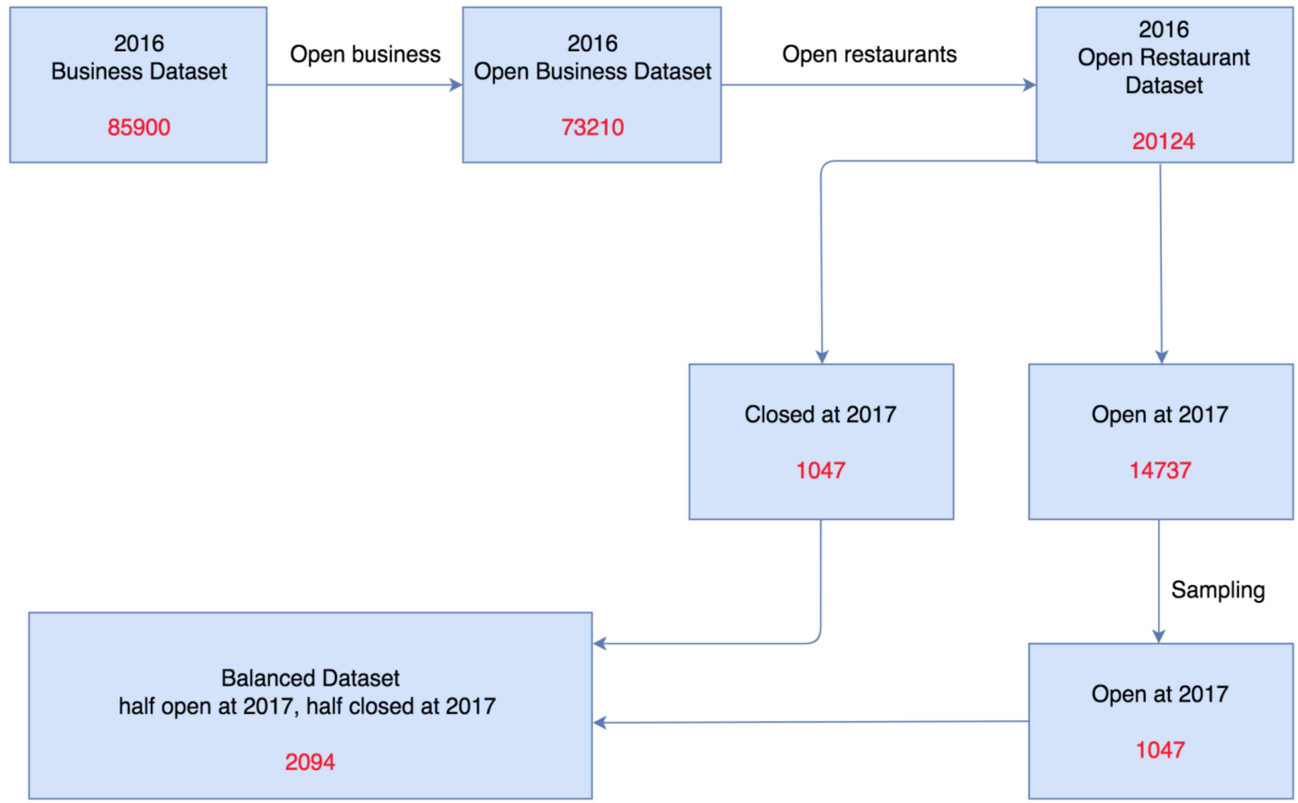
INTRODUCTION

- More and more people choose Yelp to help making daily decisions
- It would be fun to see if the future development of certain restaurants can be predicted through current data
- Might help investors make better decisions



DATASET DESCRIPTION

- Two databases with identical fields but different release time (2016,2017)
- **Aim to get restaurants closed in this one year period**



FEATURE ENGINEERING

Text Features	Unigram	Good
		Bad
	Bigram	Sanitation
		Location
		Service
		Taste
Non-text Features	Trend	Review star loss
	Business	Review count
		Chain restaurant
		Return guest count
		Restaurant type
	Location	Nearby restaurant comparison
		(City economic status)

TEXT FEATURES - Unigram (2)

- Using a sentiment dictionary to **catch** certain sentiment words
    - eg. “unigram_good”: 'love', 'nice', 'delicious', 'amazing', 'top', ’favorite’, etc.

      “unigram_bad”: 'nasty', 'noisy', 'disappoint', 'cockroach', 'fly', 'mosquito', etc.

- Count number of word occurrence for all reviews with same business
- NOTICE: only TWO features generated finally

A simple example...

restaurant name	reviewer	restaurant name	uni_good	uni_bad
Outback Steakhouse	Jack	The food here is amazing! One of my favorite restaurant in Chapel Hill. The environment is a little bit noisy however...	2	1
Burger King	Andrew	The food here is amazing! One of my favorite restaurant in Chapel Hill. The environment is a little bit noisy however...	4	2
	Sam	The food here is amazing! One of my favorite restaurant in Chapel Hill. The environment is a little bit noisy however...		

TEXT FEATURES - Bigram (8)

- Want to discover which parts are critical for business success
- Construct Bigram features by different categories
  - Sanitation (2)
  - Location (2)
  - Service (2)
  - Taste (2)
- Find co-occurrence of pair of words in each sentence

Bigram - Sanitation (2)

- “sanitation_good”
  - eg. environment...clean, atmosphere...quiet, etc.
- “sanitation_bad”
  - eg. environment...nasty, table...dirty, etc.

Another example :)

<table>
<tr><th>restaurant name</th><th>reviewer</th><th>restaurant name</th><th>sani_good</th><th>sani_bad</th></tr>
<tr><td>Outback Steakhouse</td><td>Jack</td><td>I love the atmosphere here, its just so quiet... The overall environment is clean except the table is a bit dirty.</td><td>2</td><td>1</td></tr>
<tr><td rowspan="2">Burger King</td><td>Andrew</td><td>Won't come here again! Bad environment and nasty floor. 1/5</td><td rowspan="2">0</td><td rowspan="2">3</td></tr>
<tr><td>Sam</td><td>Love the burger. The environment is bad though.</td></tr>
</table>

Bigram - Service (2)

- “Service_good”
  - eg. waiter…helpful,service...fantastic, etc.
- “Service_bad”
  - eg. waitress...worst, staff...disrespect, etc.

Bigram - Location (2)

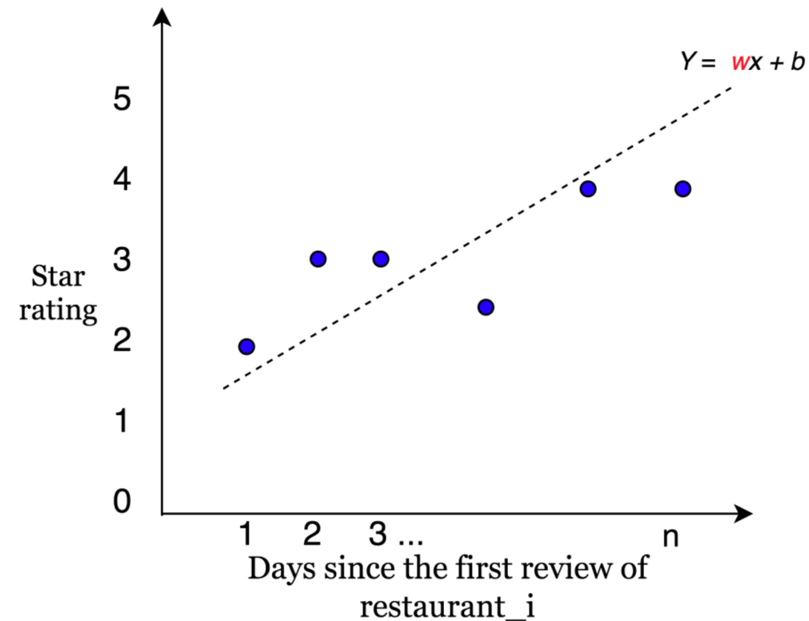
- “location_good”
  - eg. place…cool, parking...easy, etc.
- “location_bad”
  - eg. place...crowded, bar...boring, etc.

Bigram - Taste (2)

- “Taste_good”
  - eg. drink...best, dessert...wonderful, etc.
- “Taste_bad”
  - eg. food...nasty, appetizer...disgusting, etc.

NON-TEXT FEATURES (5)

- Trend
    - Star gain/loss coefficients
- Business
    - Review count
    - Chain restaurant
    - Return guest count
    - Restaurant type
- Location feature
    - Nearby restaurants comparison (not finished)
    - City economic status (failed)



Final Feature table looks like...

restaurant_id	uni_good	...	star_coeff	chain	...	Open_2017
0001	...	...	...	...	...	True
0002	...	...	...	...	...	False
0003	...	...	...	...	...	True
0004	...	...	...	...	...	True
...	...	...	...	...	...	...
2094	...	...	...	...	...	True

EXPERIMENT

- 10-fold Cross-Validation
- Logistic Regression
- Feature ablation study
- Accuracy, Precision,Recall, Precision-Recall curve

RESULT...

RESULTS

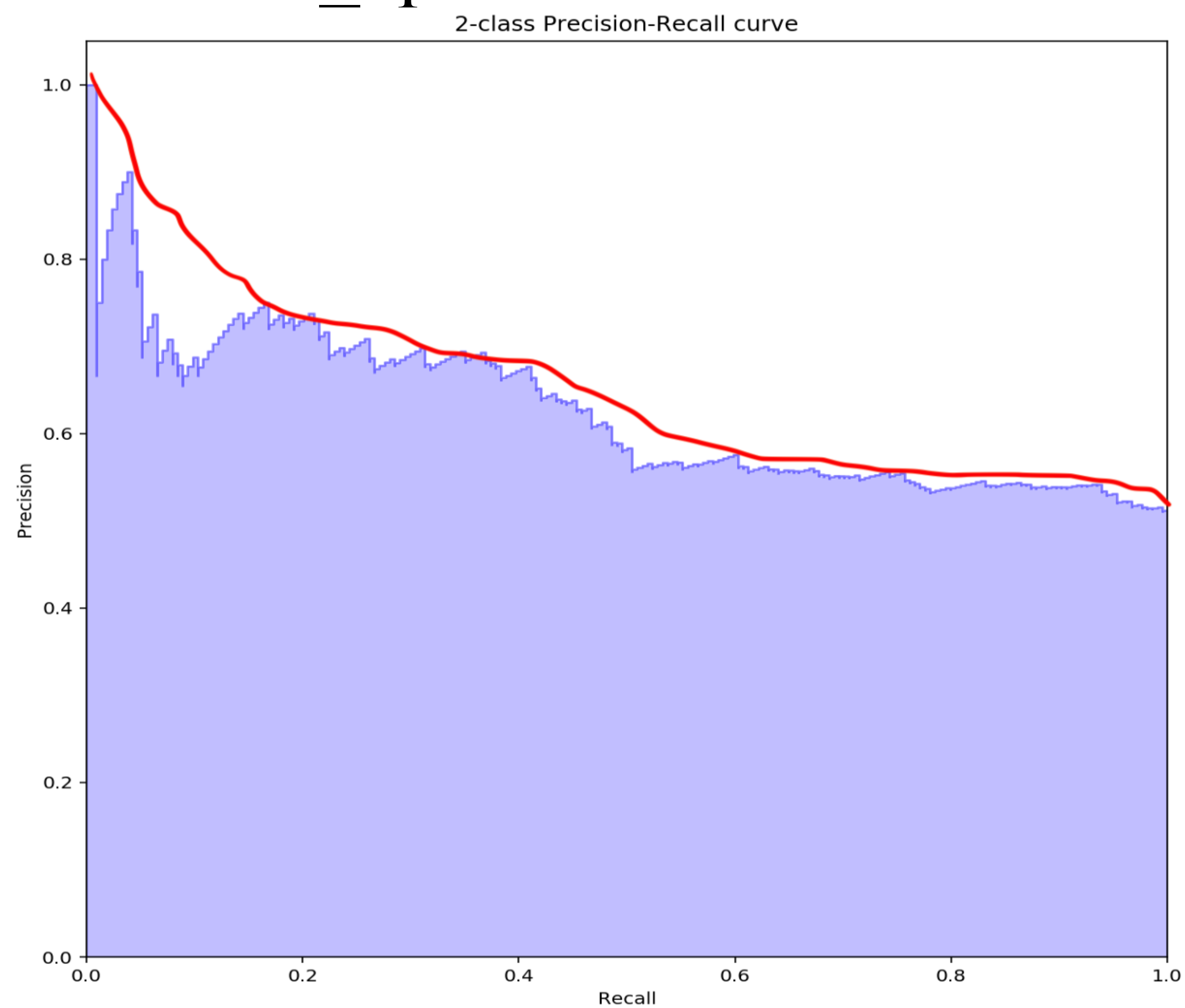
Accuracy: **62.34%**

Precision (for open): **0.696**

Recall: 0.442

		Predicted	
		Open	Closed
True	Open	462	584
	Closed	202	839

Precision - Recall curve for label_open

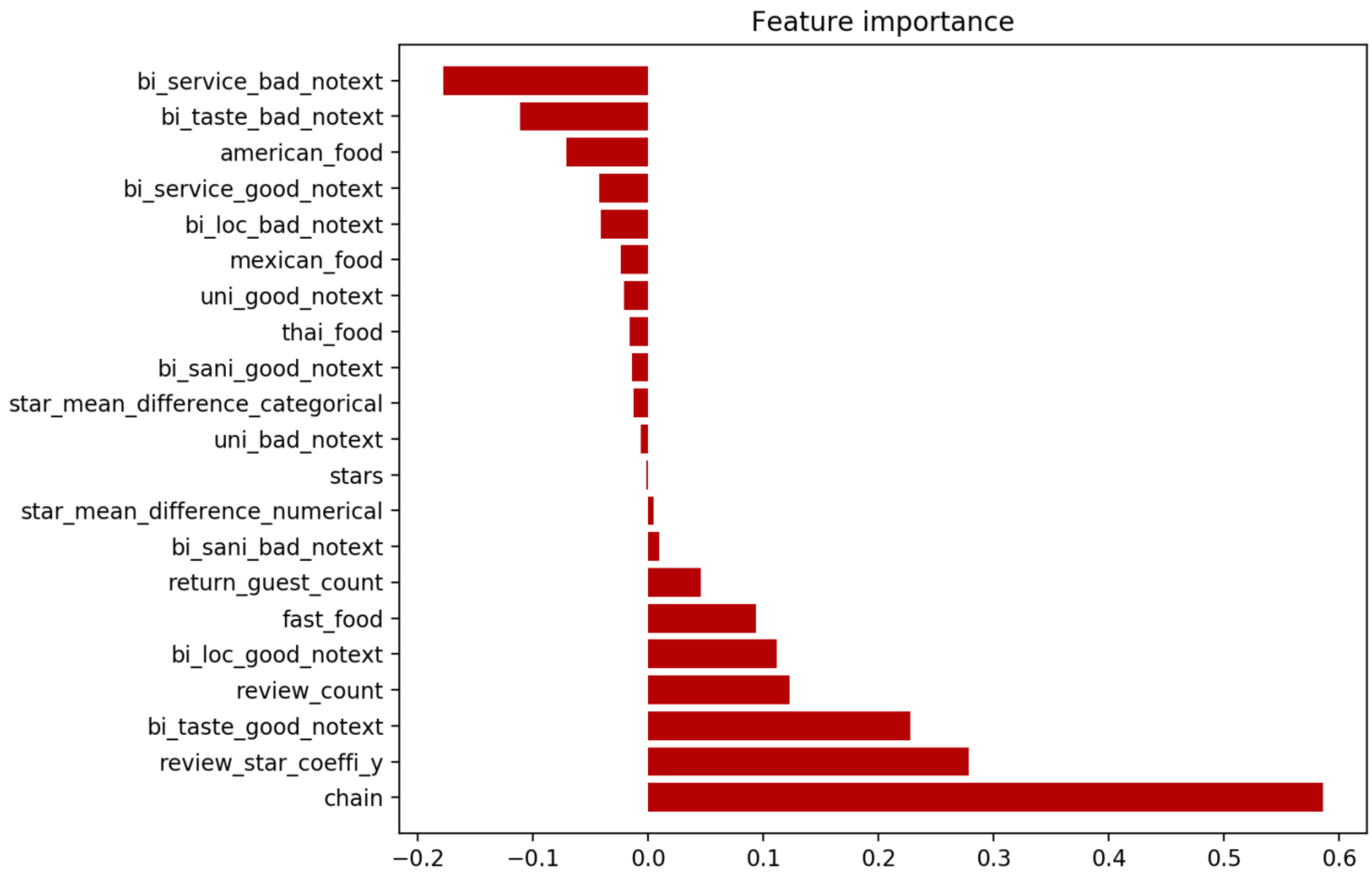


Feature ablation study

- Business features are the most importa
- Text features does not work as desired
    - Why?

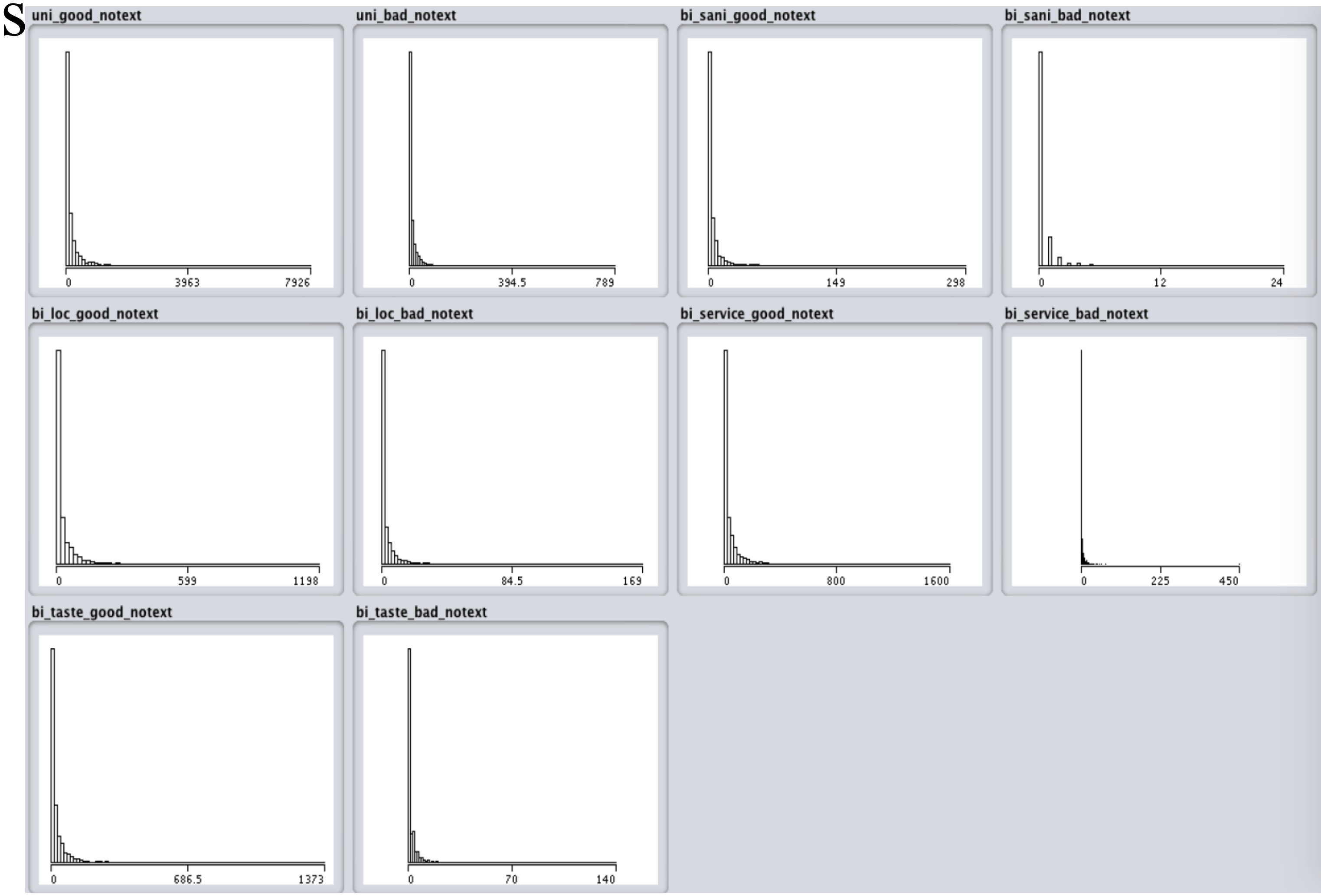
All features	Accuracy	Precision
All	0.6234	0.696
-Text features	0.6229 ▼ (-0.05%)	0.701
-Non-text features	0.5199 ▼ (-10.35%)	0.534
-Unigram	0.6243 ▲ (+0.09%)	0.696
-Bigram	0.6234	0.698
-Trend	0.6224 ▼ (-0.1%)	0.698
-Business	0.5141 ▼ (-10.93%)	0.520

Error Analysis



Error Analysis

- Too sparse
- Look back into dictionary



Error Analysis

- potential solution: Add more words
- Look back into training set and do supervised feature selection

```
sani_noun_notext_list = ['sanitation','environment','health','hygiene', \
                         'surrounding','floor','table']
sani_good_notext_list = ['clean','quiet']
sani_good_notext_list.extend(uni_good_notext_list)
sani_bad_notext_list = uni_bad_notext_list

loc_noun_notext_list = ['location','place','bar','bartenders','bartender', \
                        'atmosphere','parking','beach']
loc_good_notext_list = ['easy','pleasant','pleased','fun']
loc_good_notext_list.extend(uni_good_notext_list)
loc_bad_notext_list = ['hard','busy','annoy','underground']
loc_bad_notext_list.extend(uni_bad_notext_list)

service_noun_notext_list = ['service','quality','staff','waiter','waitress', \
                            'prepare','price']
service_good_notext_list = ['24hour','welcoming','fantastic','nice', \
                            'communicative','helpful','quick','fast','super']
service_good_notext_list.extend(uni_good_notext_list)
service_bad_notext_list = ['bad','worse','worst''confusing','improper','late', \
                           'disrespect','tragic']

taste_noun_notext_list = ['taste','food','drink','appetizer','dessert']
taste_good_notext_list = ['great','good','fantastic','pleasant','pleased','quick', \
                          'fast','decent','organic','inexpensive','cheap','fresh']
taste_bad_notext_list = uni_bad_notext_list
```

Error Analysis

- City economic status feature doesn't work
- Not all city data are released