Yelp Text Analysis

“Prediction of Restaurant Review Rating Using Review Text”

Reza Taeb
Springboard Bootcamp
Data Science Capstone Project
Spring-Summer 2018

Outline

- Objective
- Audience
- Dataset
- Findings
- Recommendations
- Suggestions for Future Research

Objective

- Intense Competition ⇨ Continuous Improvement for Survival
- “Business Reputation” is one of the major aspect that owners should always take care of and improve it.
- Reputation can be reflected in the **customer reviews** on online platforms such as **Yelp**.

“ Prediction of restaurant ratings and recognizing its building blocks through review texts can help business owners know their advantageous and disadvantageous better. “

Audience

- Current restaurant owners
- Prospective restaurant owners
- Owners of other businesses

Dataset : Source & Attributes

Source

- The dataset is provided by Yelp for its round 10th challenge (downloaded on January 9th, 2018) .
- Two sub datasets of whole dataset have been used in this project: **business, review**.

Attributes

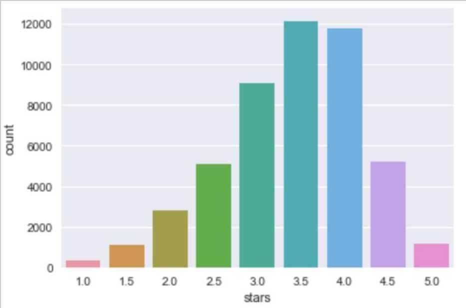
business	review
Business_id City State Postal_code Stars Review_count Categories	Business_id Stars Date Text Funny Cool Useful

Dataset : Wrangling

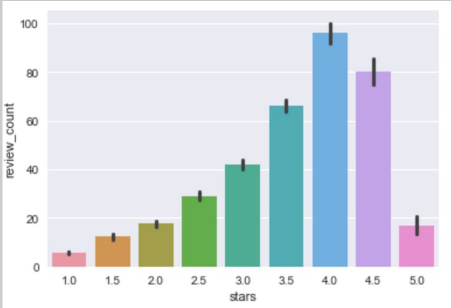
Some Data Wrangling steps that have been done on the original dataset are:

- Filtering “Restaurants” from original dataset
- Filtering “Non_English” reviews
- Adding “Country” column
- Filtering 7 most popular types of Food (Adding “food_type” column)
- Adding “word_count” and “word_list” column
- Punctuation removal
- Stop words removal
- Lemmatization

Findings



“Rating” (3.5 - 4 peak)



“Review counts” (4 - 4.5 peak)

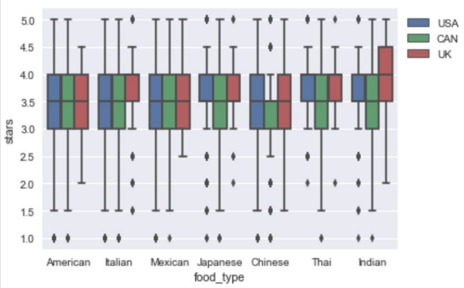
	count	mean	std	min	25%	50%	75%	max
country								
CAN	16819.0	3.426333	0.769820	1.0	3.0	3.5	4.0	5.0
UK	1549.0	3.765655	0.686447	1.0	3.5	4.0	4.0	5.0
USA	30196.0	3.395069	0.803634	1.0	3.0	3.5	4.0	5.0

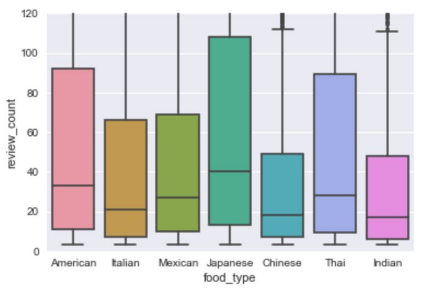
- People in **UK give better ratings** to restaurants, although American and Canadian people, respectively write **more reviews** for restaurants on Yelp.

Findings

	count	mean	std	min	25%	50%	75%	max
food_type								
American	8401.0	3.402928	0.713942	1.0	3.0	3.5	4.0	5.0
Chinese	3318.0	3.286468	0.689724	1.0	3.0	3.5	4.0	5.0
Indian	1209.0	3.579404	0.690748	1.0	3.0	3.5	4.0	5.0
Italian	3788.0	3.478221	0.773531	1.0	3.0	3.5	4.0	5.0
Japanese	1972.0	3.582657	0.662095	1.0	3.0	3.5	4.0	5.0
Mexican	3604.0	3.380827	0.768168	1.0	3.0	3.5	4.0	5.0
Thai	989.0	3.582406	0.707131	1.0	3.0	3.5	4.0	5.0

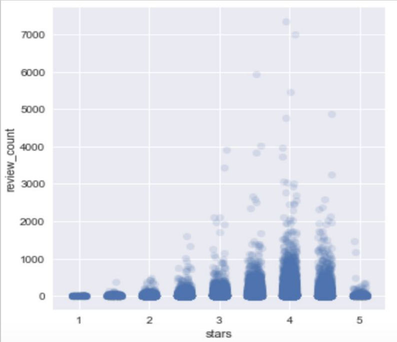
- **Type of food** can absolutely impact the **restaurants' ratings** (Japanese, Indian and Thai foods usually get higher ratings) and impact the users' **tendency to write more reviews** (Japanese food restaurants receives the highest number of reviews).

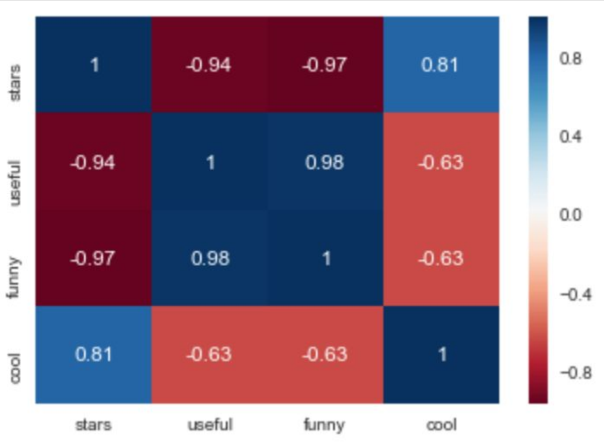




Findings

- Very **weak linear correlation** (r = 0.13) between “restaurants rating” and “number of reviews”.



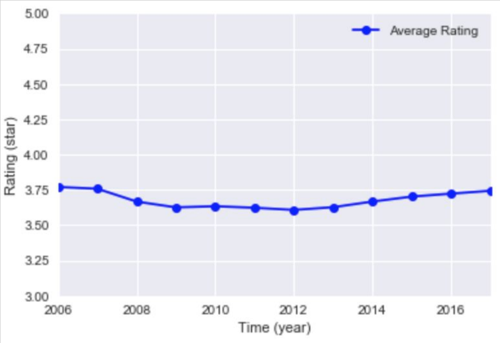


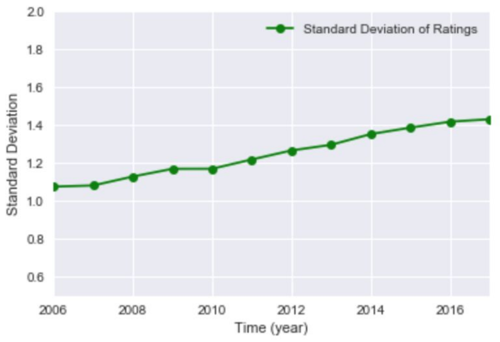
- Lower rating, higher perceived **“funny”** and **“useful”**.
- Lower rating, **longer length** (more words in review’s text).

Findings

- Through times (2006 - 2017) restaurants' ratings tend to be distributed around the same average (**around 3.65**) every year, however by passing time the rating distribution get more divers (**standard deviation rose almost 32 percent**).

“Users get more attentive”





Findings

"20 Common Words in each restaurant rating group"

1 star

	Word Freq	Percentage
food	1564	1.431251
order	1470	1.345230
get	1368	1.251887
go	1271	1.163121
place	1077	0.985587
come	1026	0.938916
time	888	0.812629
service	841	0.769618
say	837	0.765957
like	786	0.719286
us	769	0.703729
ask	755	0.690917
take	753	0.689087
one	716	0.655228
back	709	0.648822
would	700	0.640586
wait	648	0.592999
even	641	0.586593
never	607	0.555479
tell	578	0.528941

2 star

	Word Freq	Percentage
food	1446	1.439536
order	1151	1.145855
get	1000	0.995530
good	984	0.979602
go	963	0.958695
place	945	0.940776
like	884	0.880049
come	785	0.781491
time	738	0.734701
service	700	0.696871
one	597	0.594331
say	573	0.570439
would	559	0.556501
really	536	0.533604
us	525	0.522653
back	517	0.514689
try	458	0.455953
take	458	0.455953
make	457	0.454957
wait	450	0.447989

3 star

	Word Freq	Percentage
food	1987	1.351526
good	1953	1.328400
get	1467	0.997830
place	1447	0.984227
order	1395	0.948857
like	1253	0.852271
go	1238	0.842068
come	1086	0.738680
service	989	0.672702
time	939	0.638693
would	886	0.602643
really	806	0.548228
great	751	0.510818
one	746	0.507417
try	708	0.481570
make	706	0.480210
restaurant	615	0.418313
price	602	0.409471
pretty	597	0.406070
back	591	0.401989

4 star

	Word Freq	Percentage
good	3436	1.434793
food	2867	1.197192
place	2629	1.097809
get	2317	0.967525
great	2155	0.899878
go	1965	0.820538
order	1862	0.777528
come	1757	0.733682
like	1727	0.721155
service	1569	0.655178
time	1558	0.650584
try	1364	0.569575
really	1329	0.554959
one	1242	0.518630
make	1191	0.497334
also	1104	0.461005
would	1088	0.454323
nice	1071	0.447225
back	1069	0.446389
love	1026	0.428434

5 star

	Word Freq	Percentage
food	3870	1.496149
place	3489	1.348854
great	3258	1.259549
good	2572	0.994340
go	2429	0.939056
get	2279	0.881066
service	2094	0.809544
come	1966	0.760059
time	1857	0.717920
order	1847	0.714054
love	1728	0.668048
try	1649	0.637507
make	1609	0.622042
best	1586	0.613151
delicious	1555	0.601166
like	1529	0.591114
back	1488	0.575264
one	1415	0.547042
amaze	1348	0.521139
also	1210	0.467788

Springboard

yelp

Findings

N Grams (N=5)

1 star	2 star	3 star	4 star	5 star
Wanted to like this place I will never go back Would not recommend this place An hour and a half We had to ask for We will not back Would never recommend the place We will never go back We will never eat here again The quality of the food It was my first time This place used to be	I really wanted to like Wanted to like this place Out of my way to Really wanted to like this The quality of the food I felt like I was Give this place a try The food was good But We were the only ones Go out of my way I won't be going back The best part of the	Nothing to write home about The food was good but Go out of my way The food is good but The quality of the food The food was pretty good The best part of the If you are looking for Out of my way to At the end of the On the other hand was I am not a fan	This is a great place If you are looking for Is a great place to The best I've ever had I will definitely be back Just the right amount of Can't wait to go back Been here a few times I would definitely go back Get what you pay for Never had a better meal The staff is friendly	Can't wait to go back The best I've ever had I can't wait to go We will definitely go back Was one of the best I will definitely be back Can't wait to come back If you are looking for You will not be disappointed Was out of this world Just the right amount of Had in a long time

Findings

- Negative direct recommendation (exp: would never recommend) is much more common in 1 star compared to 2 star ratings.

- **"Contrast"** is very common in 3 star reviews. In this category, users usually bring both positive and negative comments in their reviews along with words such as "But" , "On the other hand" and etc.

- The obvious difference between 4 star and 5 star reviews is observing more **"excitement"** and "**exaggerated statements**" in 5 star reviews compared to 4 star reviews.

Findings

- Applying **TF-IDF technique**, we observed that words and expressions that represent feelings even in their styles (*wowowow*, *awwwwwwful*, *amaaaazing*, *looooove*) have high impact on reviews' message.

- Applying **Multinomial Naive Bayes** algorithm, just using the texts of reviews the "ratings" can be predicted in so many cases **(72 percent precision and recall for 1 star difference and 86 percent precision and recall rations for 2 star difference -between 3 and 5 stars- classification**).

Classification Report (3 and 4 stars reviews)

	precision	recall	f1-score	support
3	0.70	0.75	0.72	634
4	0.73	0.68	0.70	641
avg / total	0.71	0.71	0.71	1275

Classification Report (4 and 5 stars reviews)

	precision	recall	f1-score	support
4	0.72	0.68	0.70	1191
5	0.70	0.74	0.72	1202
avg / total	0.71	0.71	0.71	2393

Classification Report (3 and 5 stars reviews)

	precision	recall	f1-score	support
3	0.84	0.89	0.86	634
5	0.88	0.83	0.85	641
avg / total	0.86	0.86	0.86	1275

Recommendations

- **First impression** is very important for shaping users' opinions about a restaurants, therefore it is suggested to pay special attention to the first time customers.

- Excitement shows high positive correlation with perfect reviews, restaurants owners have to come up with some ideas to **boost customers excitement** in their visits.

- Still the **quality of food** is the first factor that users care about. Do not sacrifice it for anything else.

- Try to induce some suggestions in the customers' minds. "**Suggestions" (by other users and staff) can boost restaurant's reputation**. This concept may have connection with "**signature product/service"**.

Suggestions For Future Research

- Improving the mentioned classification model to get better precision and recall ratios
- Customized classification model for each country
- Customized classification model for each type of food

Links

- Full Report : https://github.com/rezataeb/Springboard/blob/master/Documents/Final-Report-RezaTaeb-%20%22Yelp%20Text%20Analysis-Prediction%20of%20Restaurant%20Review%20Rating%20Using%20Review%20Text%22%20.pdf

- Codes : https://github.com/rezataeb/Springboard/tree/master/Codes

- Dataset : https://www.yelp.com/dataset/challenge