Predict Helpful Reviews

ALI R KAYA

Table of Contents

Why Helpful Reviews?

- An outcome of a customer's experience with a product and an input for a potential customer's buying process.

- The potential buyers need to reach the most helpful customer reviews with minimum time and effort to use their resources more efficiently

- Help firms establish profitable business relationships by increasing the **likelihood of purchase**, **providing material information** about the product, or **improving customer service**.

The Dataset

- Yelp academic dataset is available for free and open to the public.

- **business.json** contains information about each company such as name and location, attributes, working hours, etc.
- **review.json** has information about each posted review such as user id, star rating, the customer review itself, number of useful votes, etc.
- **user.json** provides information about each Yelp user such as first name, the total number of reviews, the list of friends, the average star rating, etc.
- **checkin.json** has information about check-in for each business, such as business id and date.
- **tip.json** (the shorter version of reviews and conveys quick suggestions to the businesses) such as the tip itself, the number of compliments and dates, etc.
- **photo.json** contains information about each photo uploaded to Yelp, such as photo id and photo label, etc.

* It can be acquired from Yelp's official website by filling a form indicating that the it will only be used for academic and research purposes.

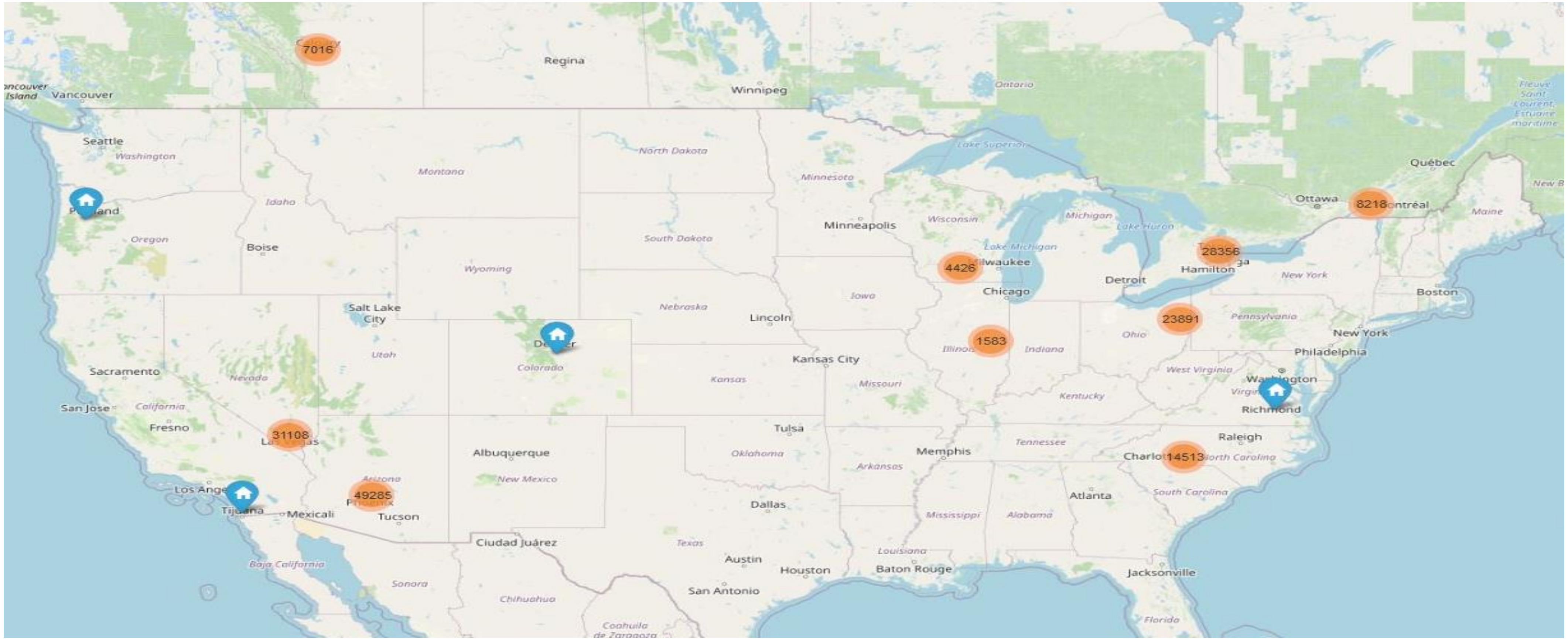
The Dataset yelp

- 8,021,122 reviews
- 1,968,703 users
- 209,393 businesses

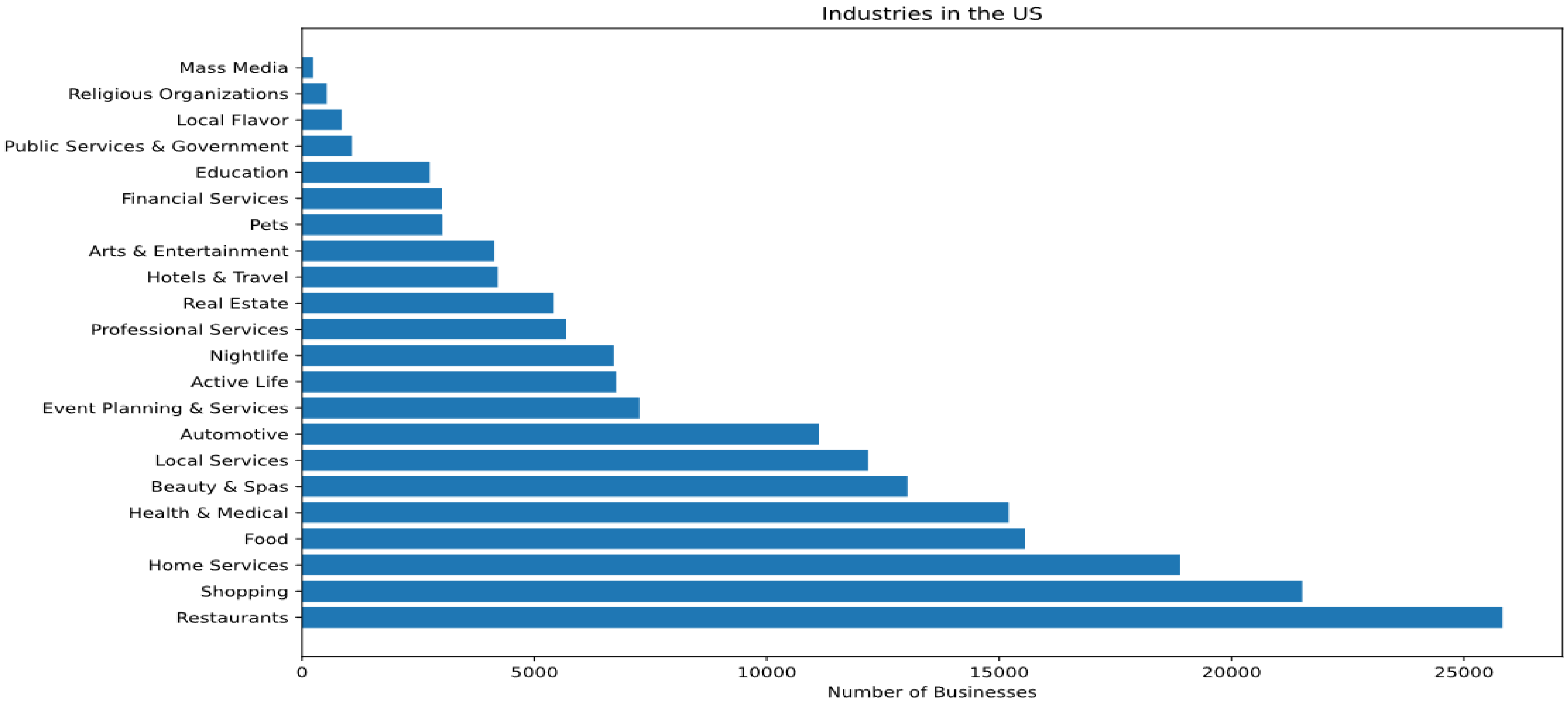
- 3.74 star rating
- 22.17 reviews
- 36.9 reviews

- 1.323 helpful vote
- 39.8 helpful vote
- 30.5% restaurants

The Dataset


7016
Regina
Winnipeg
Vancouver
Seattle
Washington
Montana
North Dakota
Minnesota
Ontario
Lake Superior
Québec
Ottawa
8218
Maine
Idaho
Oregon
Boise
South Dakota
Minneapolis
Wisconsin
Michigan
Lake Huron
Lake Michigan
4426
28356
Hamilton
Detroit
New York
Wyoming
Chicago
Boston
Salt Lake City
Nebraska
Lincoln
Iowa
23891
Pennsylvania
Ohio
1583
Indiana
New York
Philadelphia
Utah
Colorado
Kansas City
Kansas
Missouri
West Virginia
Kentucky
Sacramento
Nevada
San Jose
California
Fresno
31108
Tulsa
Raleigh
Albuquerque
Oklahoma
Arkansas
Memphis
Tennessee
14513
Arizona
New Mexico
49285
Atlanta
South Carolina
Dallas
Mexicali
Tucson
Mississippi
Alabama
Ciudad Juárez
Texas
Louisiana
Austin
Houston
Baton Rouge
Jacksonville
San Antonio
Baja California
Sonora
Chihuahua
Coahuila
Florida

The Dataset



Industries in the US
Mass Media
Religious Organizations
Local Flavor
Public Services & Government
Education
Financial Services
Pets
Arts & Entertainment
Hotels & Travel
Real Estate
Professional Services
Nightlife
Active Life
Event Planning & Services
Automotive
Local Services
Beauty & Spas
Health & Medical
Food
Home Services
Shopping
Restaurants
0
5000
10000
15000
20000
25000
Number of Businesses

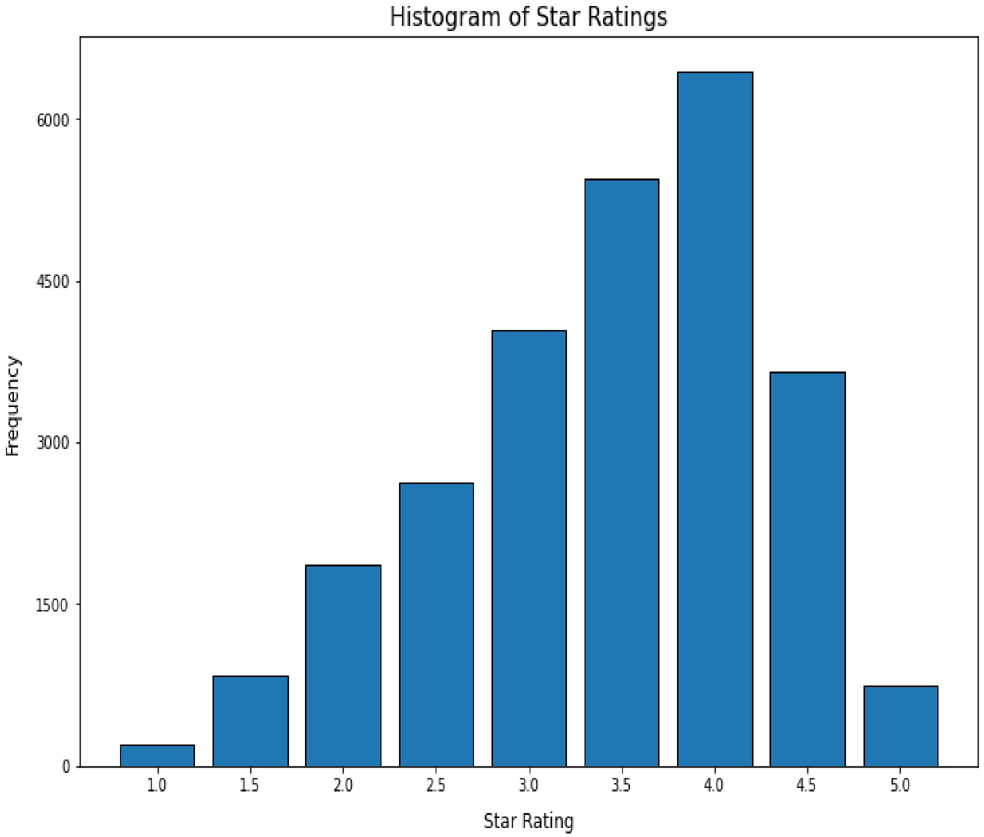
The Dataset

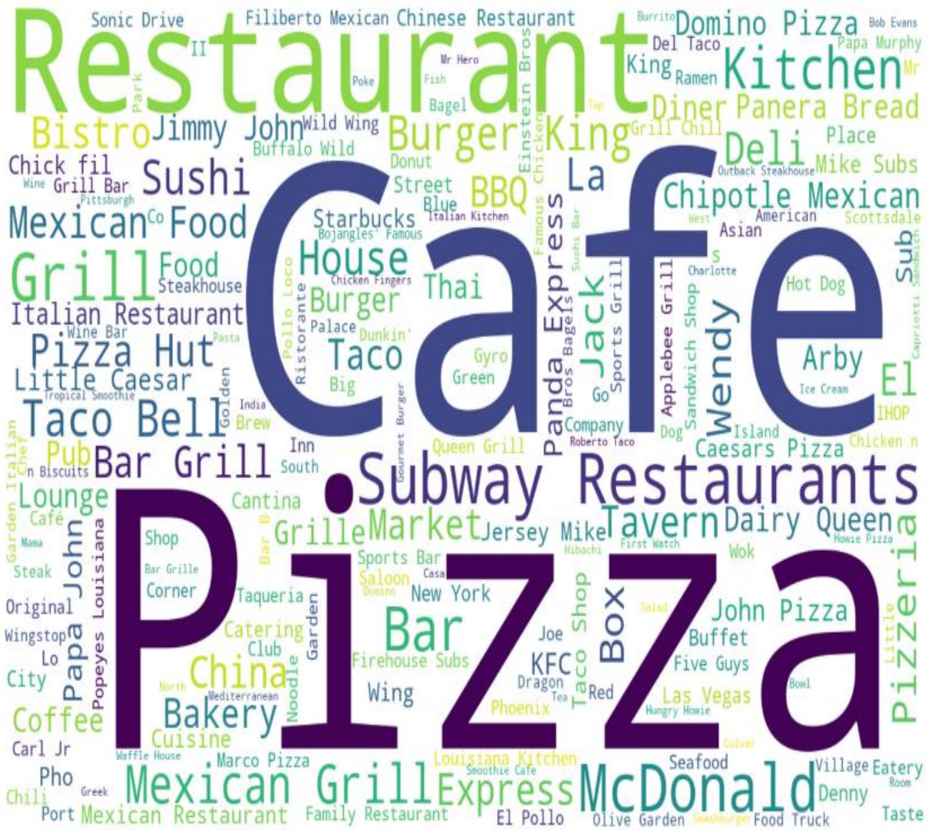
Business	Number of Branches
Subway Restaurants	609
McDonald's	536
Taco Bell	294
Burger King	273
Wendy's	232
Pizza Hut	232
Jack in the Box	182
Chipotle Mexican Grill	168
Jimmy John’s	157
Panda Express	145

- 25,827 restaurants
  - 130.41 (average) reviews
  - 10,129 (max) reviews

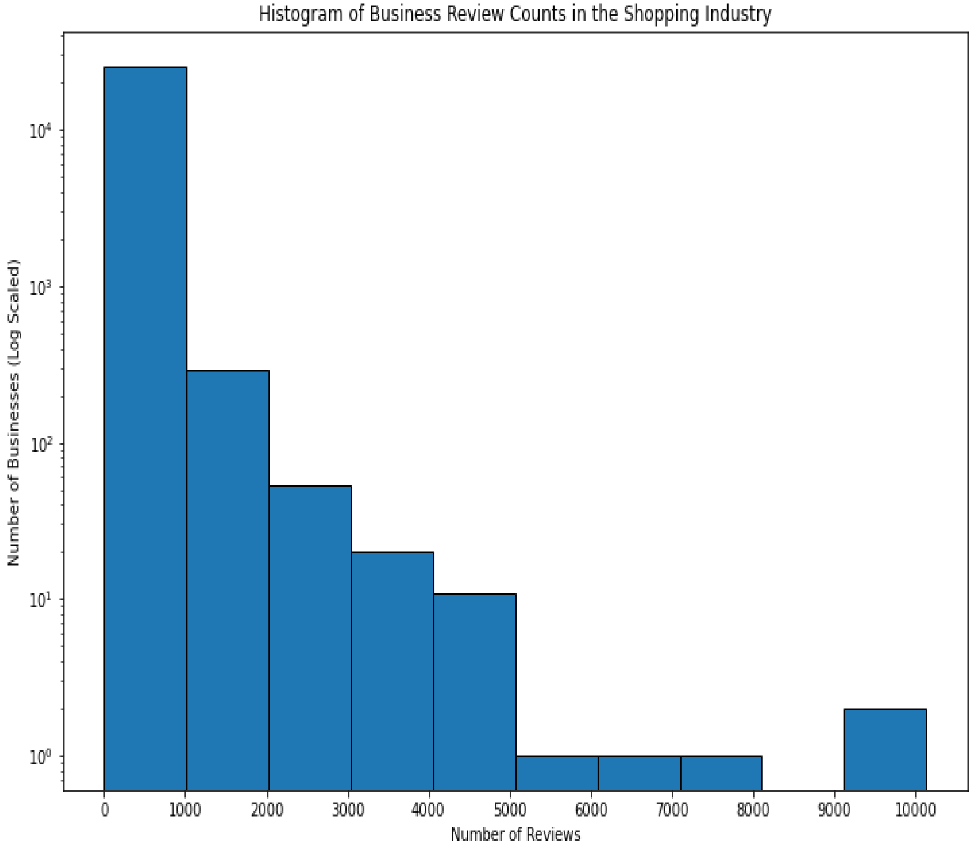
- 3,487,937 reviews
  - 1.046 (average) helpful vote
  - 758 (max) helpful vote

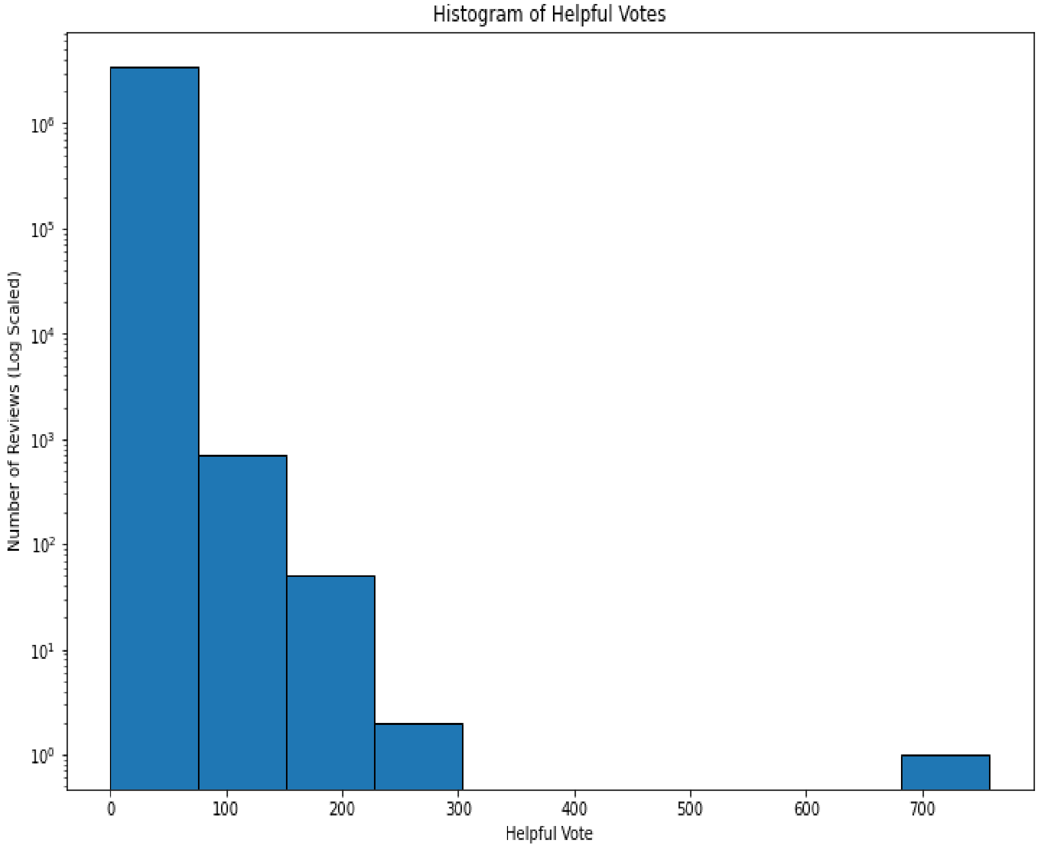
The Dataset


Histogram of Star Ratings
Frequency
0
1500
3000
4500
6000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Star Rating

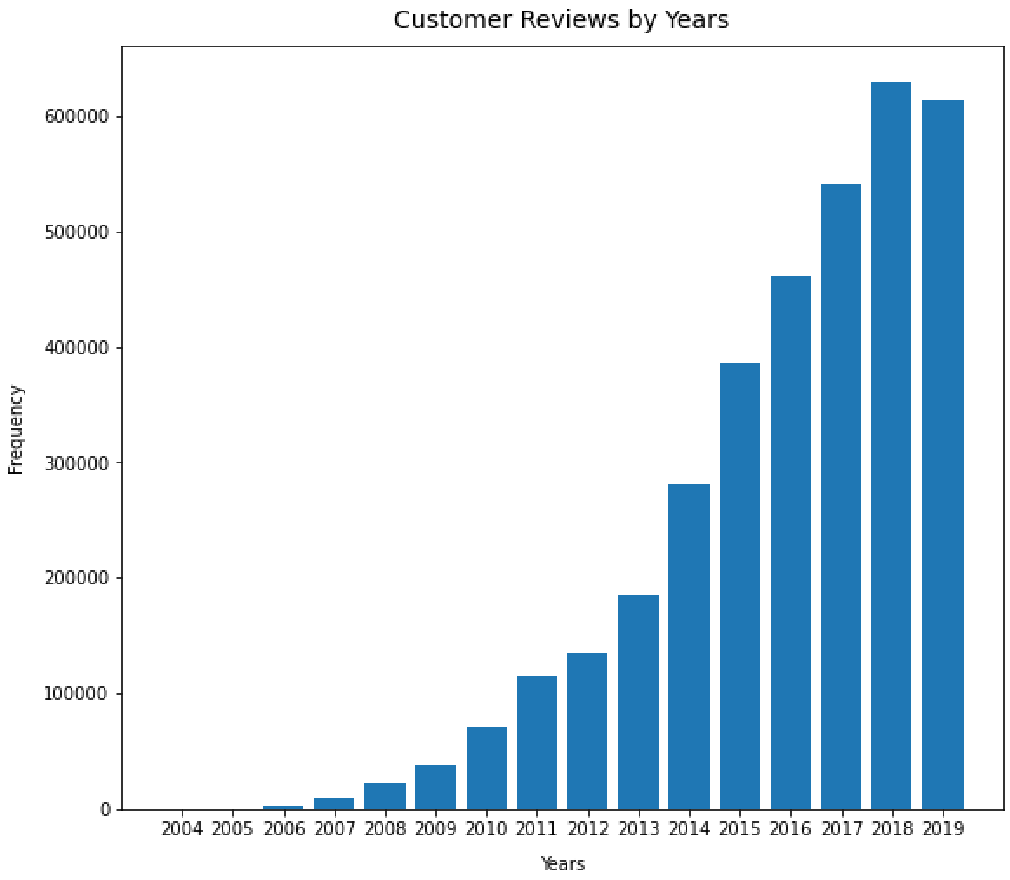

Restaurant
Cafe
Pizza
Kitchen
Subway Restaurants
Burger King
Domino Pizza
Panera Bread
Chipotle Mexican
Mexican Food
Grill
Italian Restaurant
Pizza Hut
Little Caesar
Taco Bell
Bar Grill
Sushi
Deli
House
Taco
Thai
Jimmy John
Bistro
McDonald
Mexican Grill
Express
Market
Tavern
Dairy Queen
Bakery
China
Coffee
Bar
Wendy
Panda Express
Arby
John Pizza
Mexican Restaurant

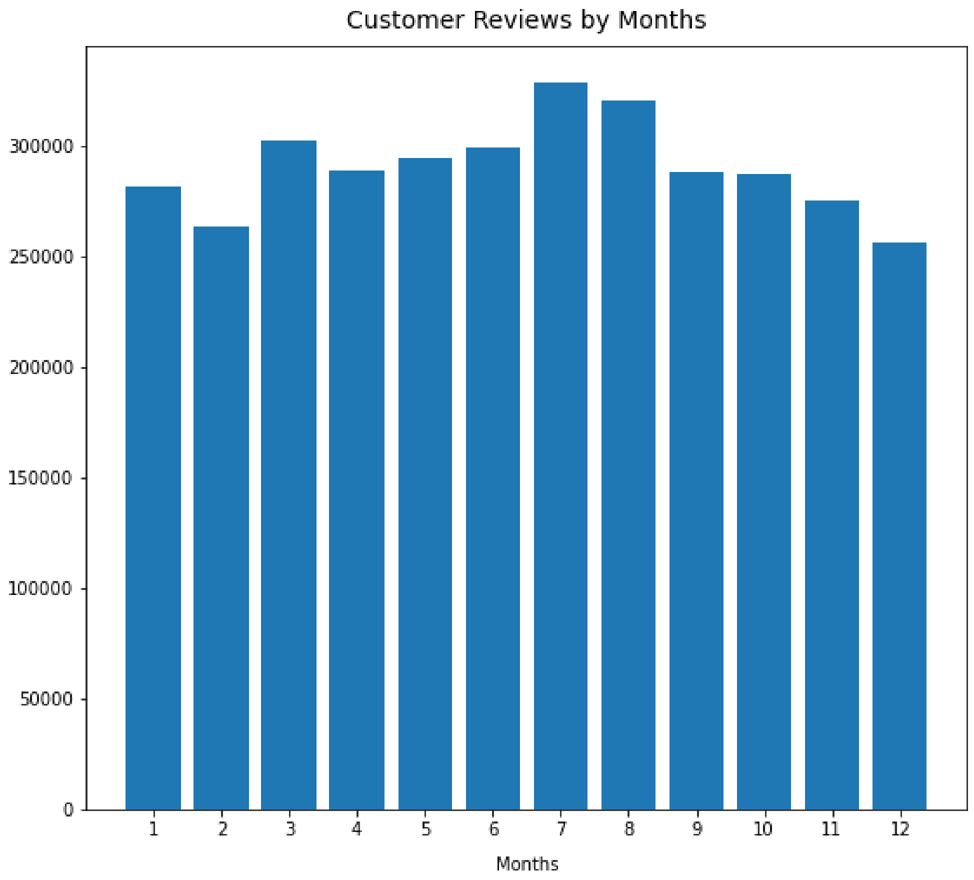
The Dataset


Histogram of Business Review Counts in the Shopping Industry
Number of Businesses (Log Scaled)
Number of Reviews
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

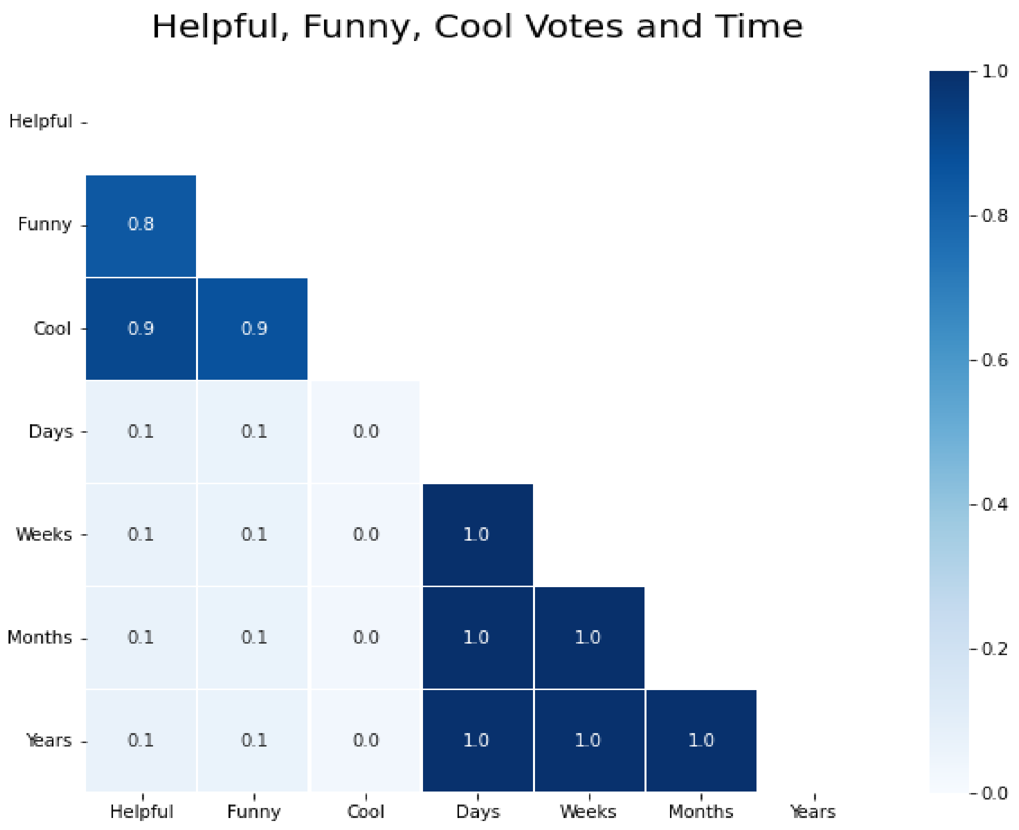

Histogram of Helpful Votes
Number of Reviews (Log Scaled)
Helpful Vote
0
100
200
300
400
500
600
700

The Dataset

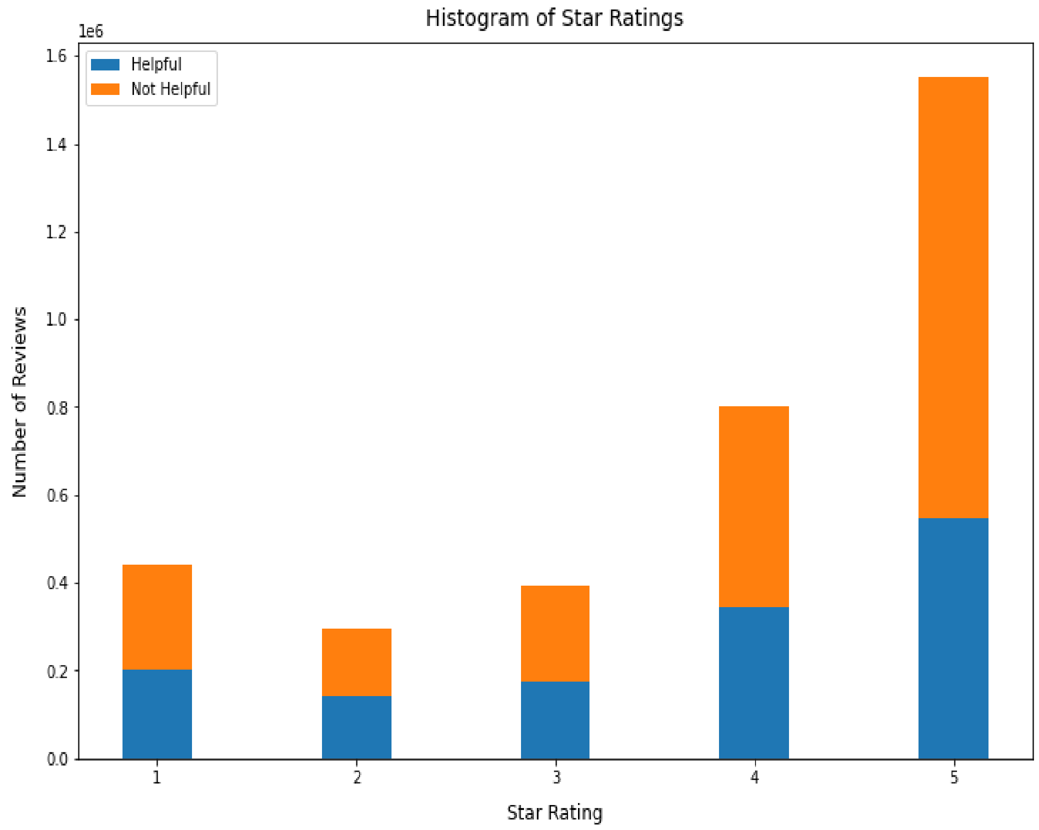

Customer Reviews by Years
Frequency
600000
500000
400000
300000
200000
100000
0
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Years


Customer Reviews by Months
300000
250000
200000
150000
100000
50000
0
1 2 3 4 5 6 7 8 9 10 11 12
Months

The Dataset



Helpful, Funny, Cool Votes and Time
Helpful
Funny
Cool
Days
Weeks
Months
Years
0.8
0.9
0.9
0.1
0.1
0.0
0.1
0.1
0.0
1.0
0.1
0.1
0.0
1.0
1.0
0.1
0.1
0.0
1.0
1.0
1.0
Helpful
Funny
Cool
Days
Weeks
Months
Years
1.0
0.8
0.6
0.4
0.2
0.0



Histogram of Star Ratings
1e6
Helpful
Not Helpful
Number of Reviews
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Star Rating

The Dataset

Most Common Phrases in 1-Star Ratings

Most Common Phrases in 5-Star Ratings

The Dataset

Name	Member Since	How Many Times Elite?	Average Star Rating	Number of Fans	Number of Reviews
Brad	2009	0	3.11	77	1259
Stefany	2011	7	3.39	785	1166
Michael	2008	7	3.90	1090	915
Karen	2006	6	3.88	479	832
Norm	2008	9	3.75	319	815
Jennifer	2010	7	3.61	98	810
Jennifer	2009	9	4.05	185	682
Deni	2010	5	3.62	154	639
Pepper	2011	0	3.35	110	626
DJ	2010	2	3.65	121	599

The Dataset

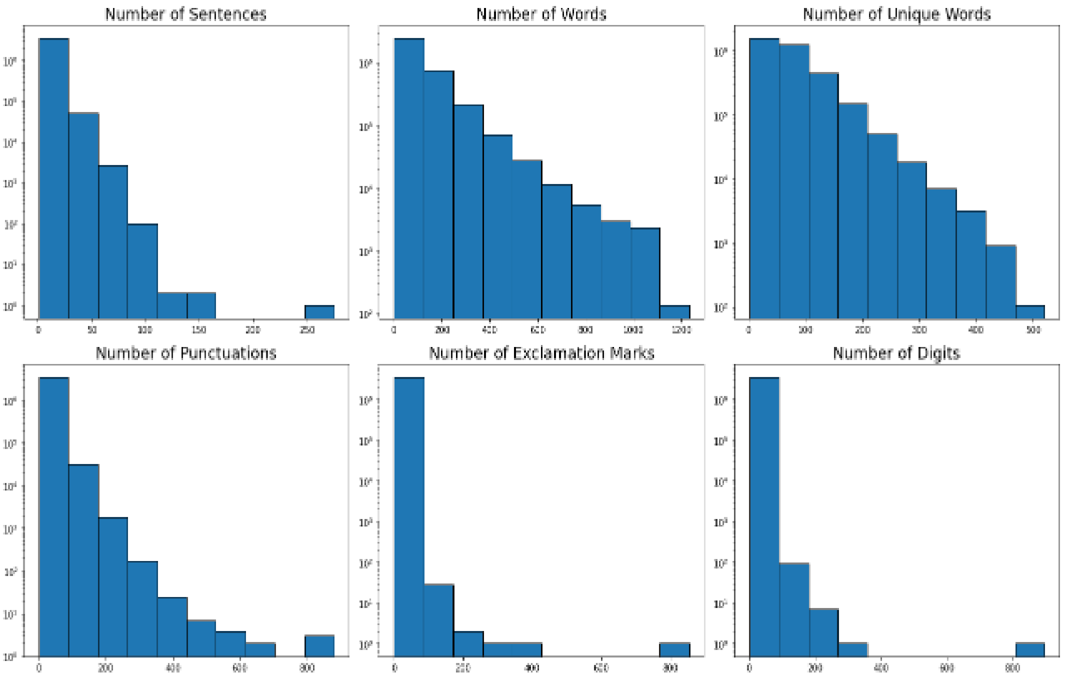
Business	City	State	Average Star Rating	Number of Reviews
Bacchanal Buffet	Las Vegas	NV	4.0	10,417
Mon Ami Gabi	Las Vegas	NV	4.0	9,536
Wicked Spoon	Las Vegas	NV	3.5	7,594
Hash House A Go Go	Las Vegas	NV	4.0	6,859
Earl of Sandwich	Las Vegas	NV	4.5	5,370
Yardbird Southern Table & Bar	Las Vegas	NV	4.5	4,979
The Cosmopolitan of Las Vegas	Las Vegas	NV	4.0	4,973
The Buffet At Wynn	Las Vegas	NV	3.5	4,953
Secret Pizza	Las Vegas	NV	4.0	4,882
Luxor Hotel and Casino Las Vegas	Las Vegas	NV	2.5	4,819

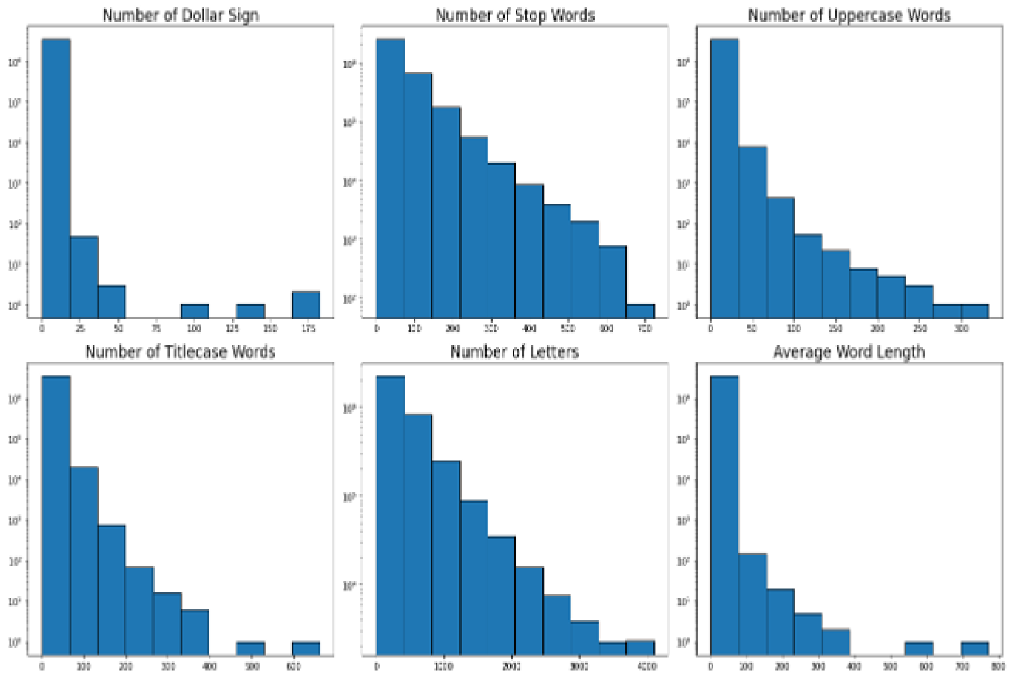
Data Cleaning & Feature Extraction

- Basic features for EDA and predictive purposes

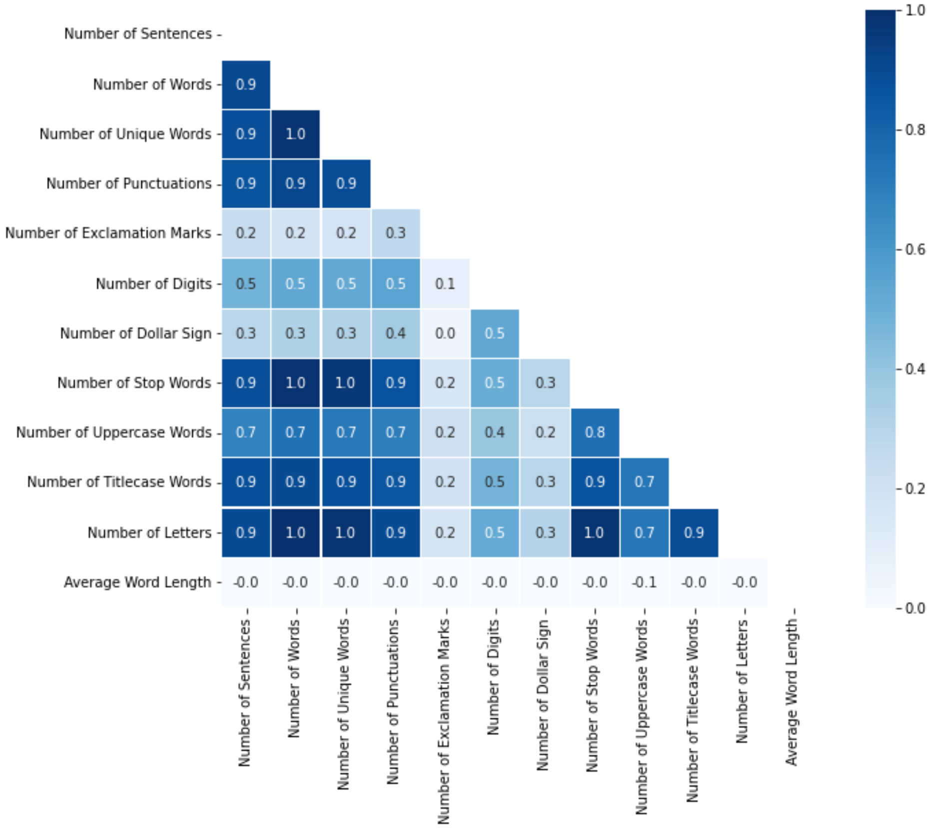
- Number of Sentences
- Number of Words
- Number of Unique Words
- Number of Punctuations
- Number of Exclamation Marks
- Number of Digits
- Number of Dollar Sign
- Number of Stop Words
- Number of Uppercase Words
- Number of Titlecase Words
- Number of Letters
- Average Word Length

Data Cleaning & Feature Extraction





Data Cleaning & Feature Extraction



	review	average_length_words
11146	フェニックス(地元の発音では、フィーニックス)のスカイハーバー国際空港の４番ターミナルにあり...	25.625
19091	ここもレストランではなく、先にカウンターで注文するファストフード店です。\nフォーなどが食べ...	48.800
20006	南部的なフライドチキンが楽しめる全国チェーン。チキンのからあげ等が大好きな日本人にとって、思...	124.000
26676	店員さんがフレンドリー！ \n味も美味しくディナーと次の日のランチも来てしまいました\nディナ...	21.000
32981	好吃好吃好吃好吃好吃好吃好吃好吃好吃今天和我男朋友一起来吃面，好吃好吃好吃好吃我写了那么长了...	92.000

Feature	Correlation with Helpful Votes
Number of Sentences	0.266259
Number of Words	0.281358
Number of Unique Words	0.287694
Number of Punctuations	0.280379
Number of Exclamation Marks	0.066640
Number of Digits	0.164860
Number of Dollar Sign	0.123075
Number of Stop Words	0.261895
Number of Uppercase Words	0.207940
Number of Titlecase Words	0.294231
Number of Letters	0.282718
Average Word Length	-0.018440

Data Cleaning & Feature Extraction

- Feature extraction for predictive modelling

- Number of Photos
- Number of URLs
- Number of Price
- Number of Time
- Number of Emoticons

Feature	Pearson R	Significance
PHOTO	0.07	0.00
URL	0.05	0.00
PRICE	0.13	0.00
TIME	0.07	0.00
EMOTICON	0.16	0.00

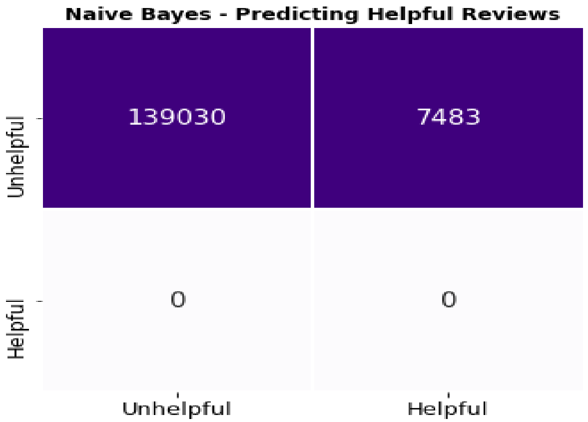
Data Cleaning & Feature Extraction

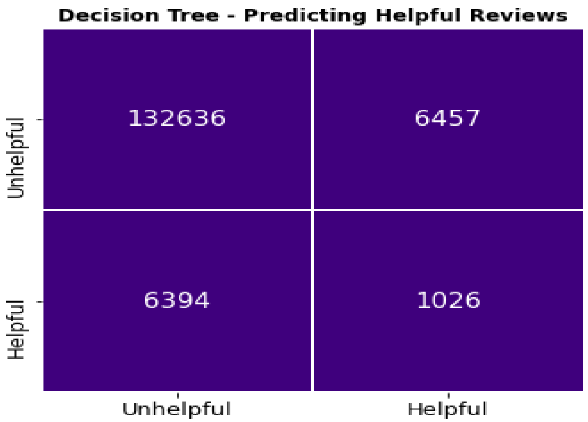
- Data Cleaning Steps
  - Replace Chinese and Japanese characters with whitespace
  - Whitespace formatting
  - Reduce duplicated letters (Ex. Sooooooooooooooooo → So)
  - Replace spaced words (Ex. A M A Z I N G → AMAZING)
  - Fix contractions (Ex. I’m → I am)
  - Remove hashtags (#) and mentions (@)
  - Remove punctuations
  - Remove digits
  - Lowercase terms
  - Remove stop words
  - Lemmatize and
  - Stemmer

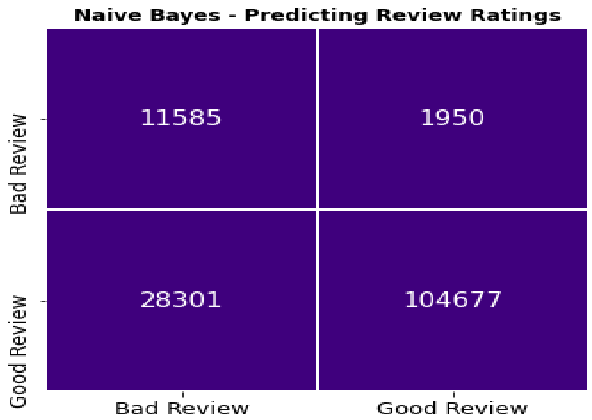
Data Cleaning & Feature Extraction

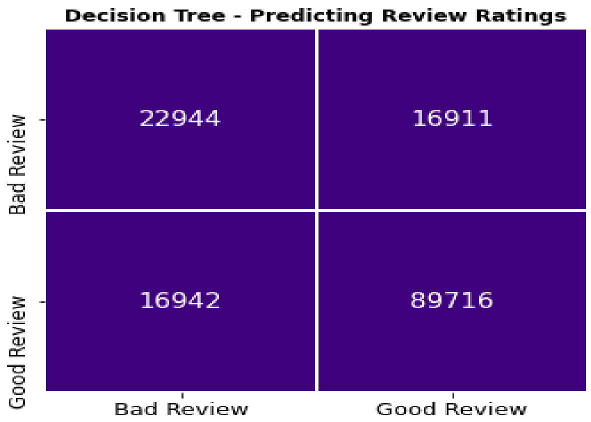
Term	Frequency	Term	Frequency
food	535,503	**love**	195,117
good	455,635	**wait**	194,598
place	437,241	**restaur**	193,228
great	383,019	**eat**	187,496
time	314,294	**friend**	182,940
order	312,467	**amaz**	152,153
servic	308,846	**delici**	150,732
make	228,983	**nice**	148,213
back	218,896	**tabl**	140,939
vega	196,308	**drink**	138,937

Predictive Modelling





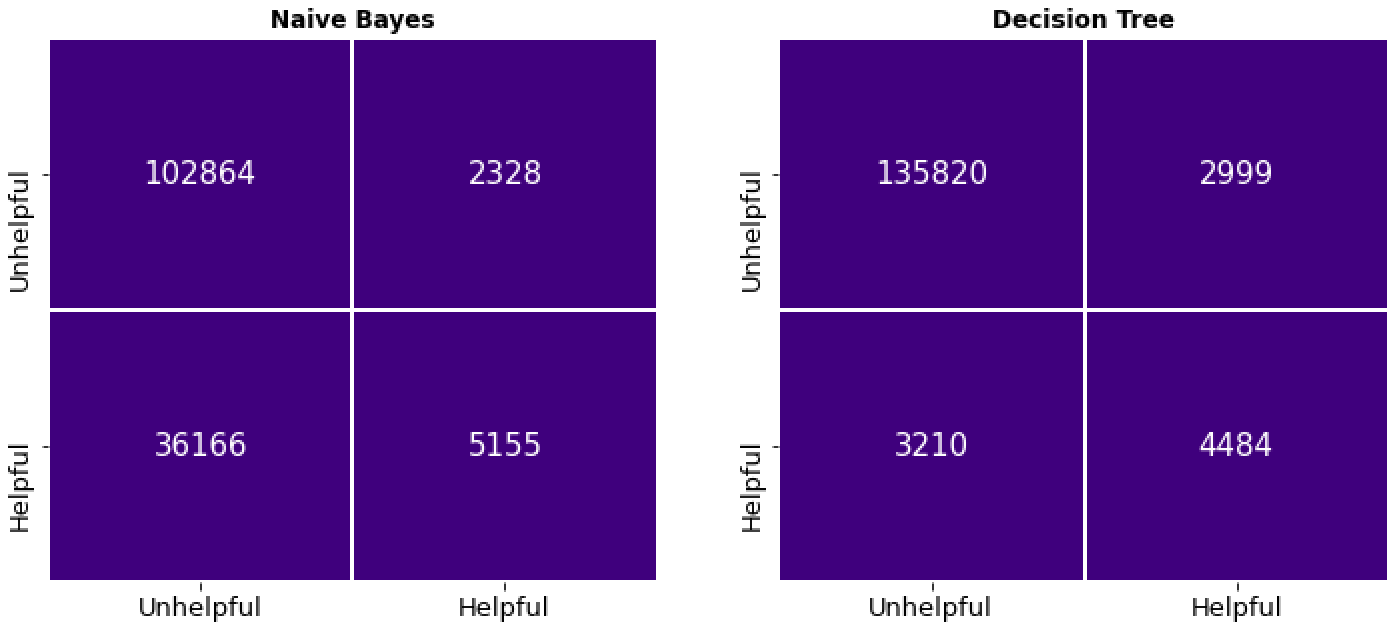




TF-IDF matrix can predict star rating but not helpful votes

Change the Features and give it another try ☺

Predictive Modelling



Model performances improved significantly after using extracted features not the TF-IDF matrix

Predictive Modelling

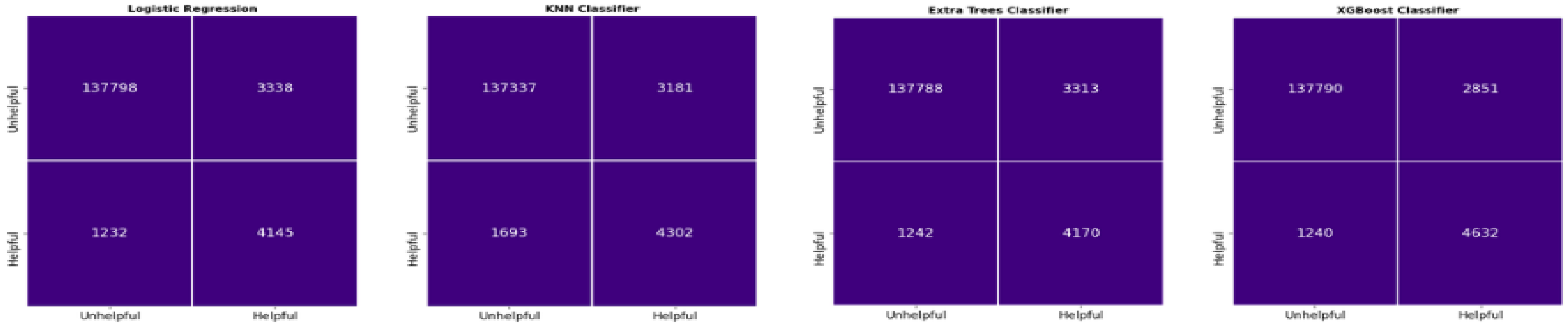
	Matthews CC	ROC Score	PR Score
Logistic Regression	0.620788	0.974573	0.723073
KNN Classifier	0.709184	0.989005	0.845982
Decision Tree Classifier	0.999946	1.000000	1.000000
Random Forest Classifier	0.999750	1.000000	1.000000
Extra Trees Classifier	0.999946	1.000000	1.000000
XGBoost Classifier	0.703560	0.983709	0.813996

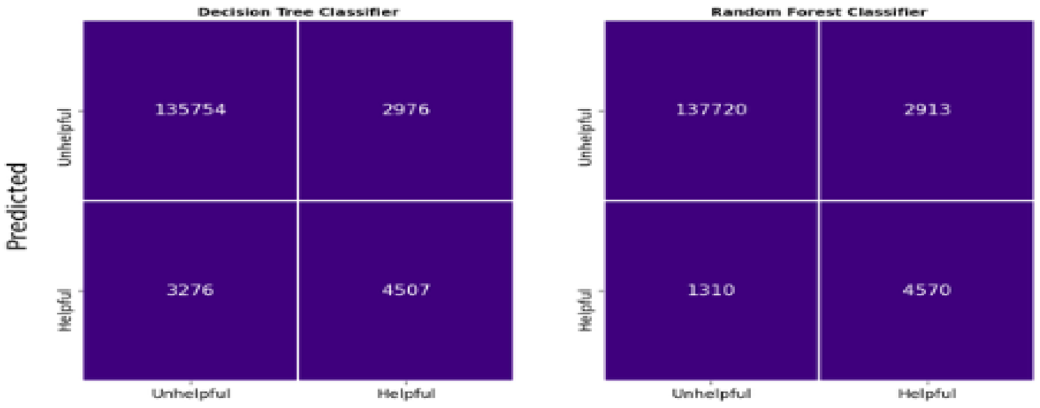
	Matthews CC	ROC Score	PR Score
Logistic Regression	0.638187	0.975003	0.727144
KNN Classifier	0.625358	0.903029	0.677878
Decision Tree Classifier	0.568085	0.789368	0.600847
Random Forest Classifier	0.674447	0.970229	0.757467
Extra Trees Classifier	0.640015	0.964758	0.724107
XGBoost Classifier	0.684751	0.980687	0.785370

Some algorithms are prone to overfitting

Let’s look at the confusion matrices

Predictive Modelling



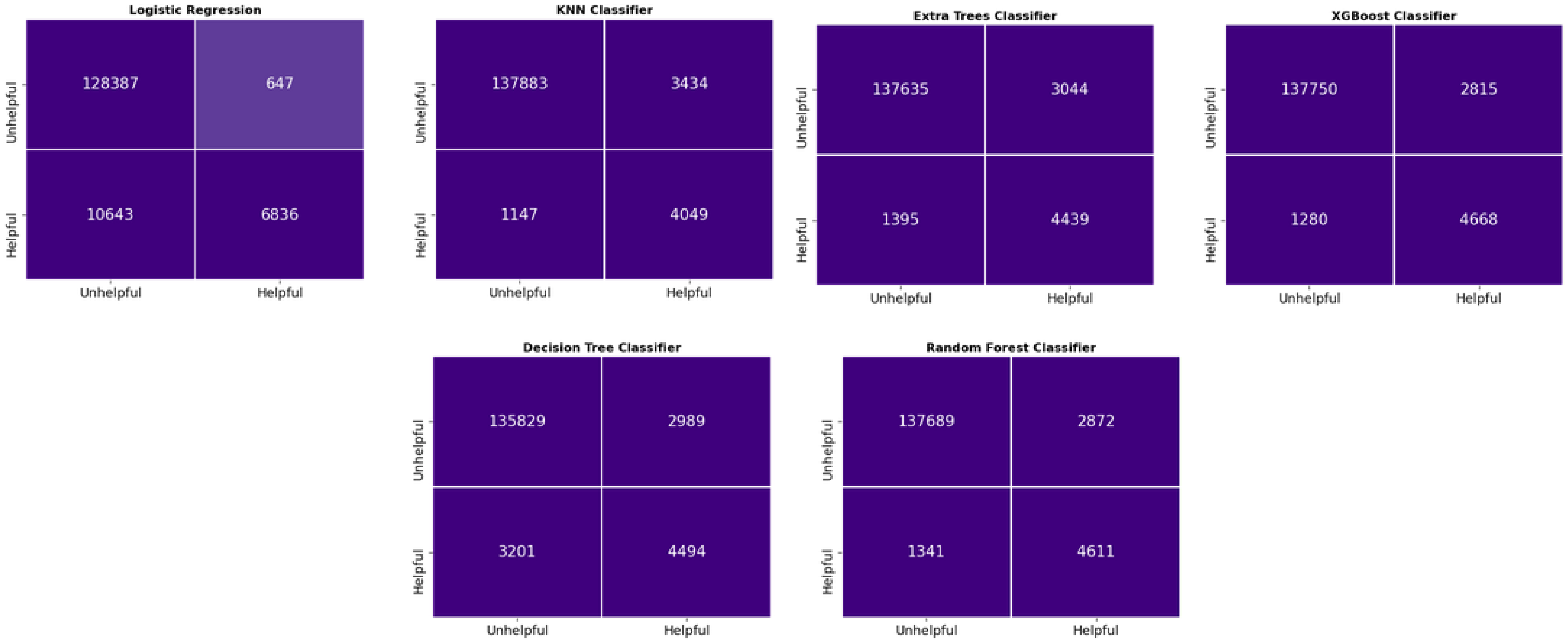


According to the test scores and the confusion matrices, we can say that the best algorithm for detecting helpful reviews is XGBoost, which has the lowest False Positive and False Negative and the highest True Positive scores.

Predictive Modelling

	Score (Default Parameters)	Score (Optimized Parameters)	Change
Logistic Regression	0.975	0.976	+ 0.001
KNN Classifier	0.897	0.950	+ 0.053
Decision Tree Classifier	0.785	0.785	0.000
Random Forest Classifier	0.969	0.976	+ 0.007
Extra Trees Classifier	0.965	0.972	+ 0.007
XGBoost Classifier	0.980	0.981	+ 0.001

Predictive Modelling



Predictive Modelling

In the confusion matrices, we see that:

- Logistic Regression predicted the highest number of helpful reviews at the expense of false positives. Moreover, it has the lowest number of false negatives among the algorithms.
- KNN predicted an average number of helpful reviews with the lowest false positive rate. However, it has the most significant number of false negatives.
- All other algorithms stay in the spectrum where the edges are Logistic Regression and KNN algorithms.

Predictive Modelling

	Recall Rate	Predicted Value		True Value
Logistic Regression	71.98 %	4,853	out of	6,742
KNN Classifier	95.16 %	904	out of	950
Decision Tree Classifier	58.40 %	4,497	out of	7,700
Random Forest Classifier	96.86 %	924	out of	954
Extra Trees Classifier	95.65 %	593	out of	620
XGBoost Classifier	97.91 %	656	out of	670

Predictive Modelling

	Recall Rate	Predicted Value		True Value
Logistic Regression	100.00 %	11	out of	11
KNN Classifier	100.00 %	11	out of	11
Decision Tree Classifier	100.00 %	11	out of	11
Random Forest Classifier	72.73 %	8	out of	11
Extra Trees Classifier	81.82 %	9	out of	11
XGBoost Classifier	72.73 %	8	out of	11

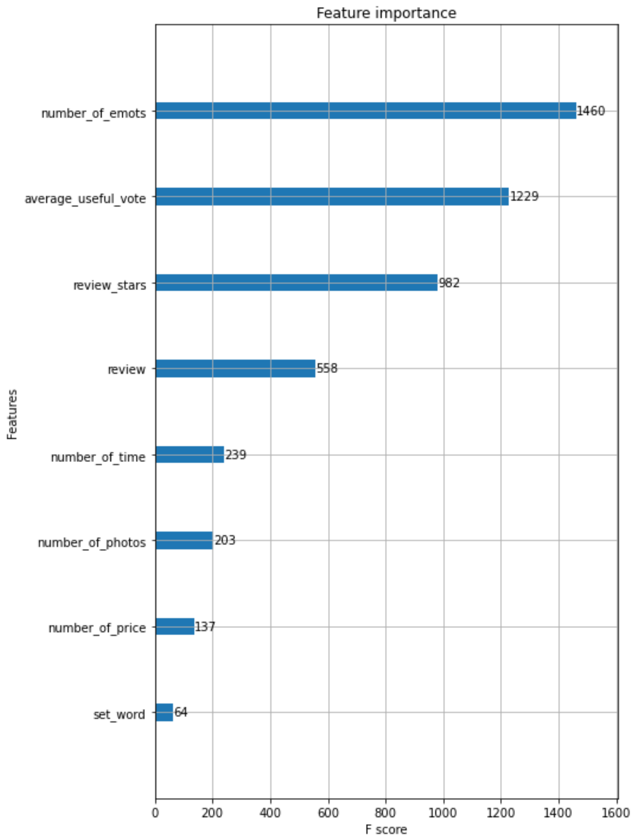
Predictive Modelling

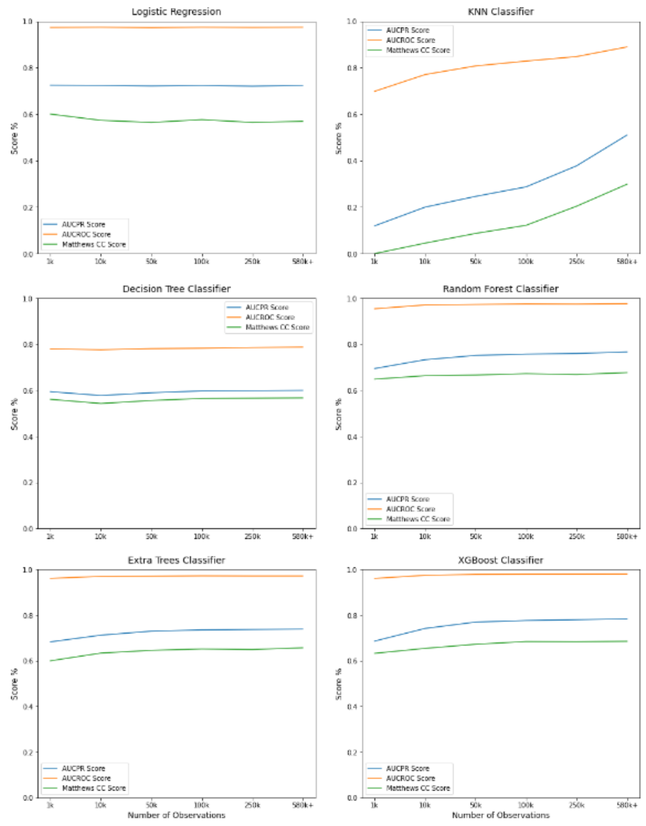
Based on the algorithms' performance in the confusion matrices and the top 5% predicted reviews, we can say that KNN is the most practical algorithm. Even though XGBoost has the best performing results, it has some flaws:

- KNN hits a 100% recall rate for the top 10 helpful reviews, but XGBoost stays at 72.73%.
- Even though KNN has a lower recall rate in general, it has the most significant number of correctly predicted helpful reviews.
- KNN provides a broader pool of helpful reviews for the business owner to hand-pick if necessary.

For those reasons, we believe that KNN is the best algorithm for our purpose in this project. We will provide some examples in the next chapters.

Predictive Modelling





Number of Emoticons is the most important features to decide if a review is a helpful review

KNN shows the most significant improvement as it fed with more data