GR5067 Final Project

A Natural Language Processing Analysis of Yelp Reviews during COVID-19

Group 15 : Samuel Frederick, Yanji Du, Junjie Ma, Yixuan Li, Xiaojia Liu

INTRODUCTION

Background:

- During the pandemic, the restaurants community was hit most severely and forced to adapt to new protocols and business mode.

Research Questions:

- We seek to understand how customers reacted to the prevalence of COVID, and how they thought about their experiences at restaurants: both sentiment and topics discussed.
- How can restaurants successfully mitigate any negative impacts of the pandemic on customer sentiment?

Main Dataset: Customer reviews of restaurants(Yelp, 2021)

Main Approach:

- *Latent Dirichlet Allocation (LDA) to* recover the most coherent topics from the corpus of reviews
- *Sentiment Analysis* of the Yelp reviews which showed suggestive evidence that more COVID cases in a county led to more negative sentiment in reviews of restaurants in that county.

DATA — Source and Processing

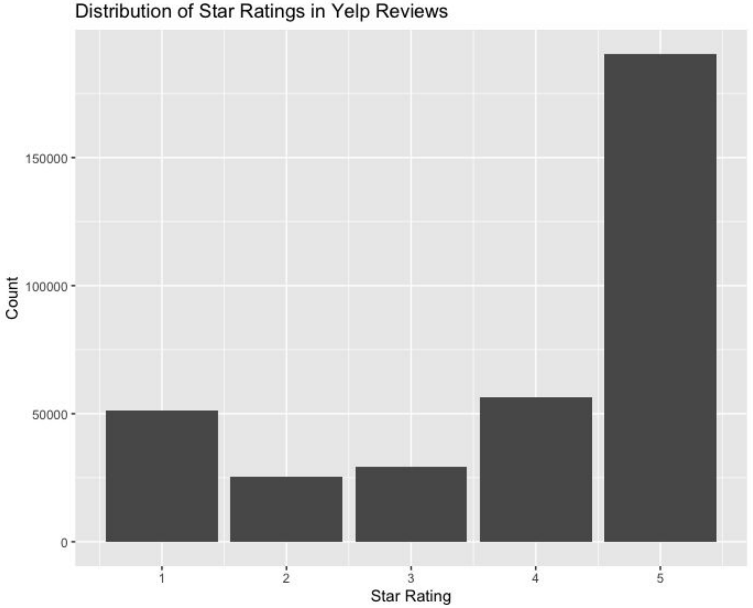
Source :

- Main: Academic dataset of Yelp business reviews (Yelp, 2021)
- Centers for Disease Control and Prevention (CDC)
- The New York Times

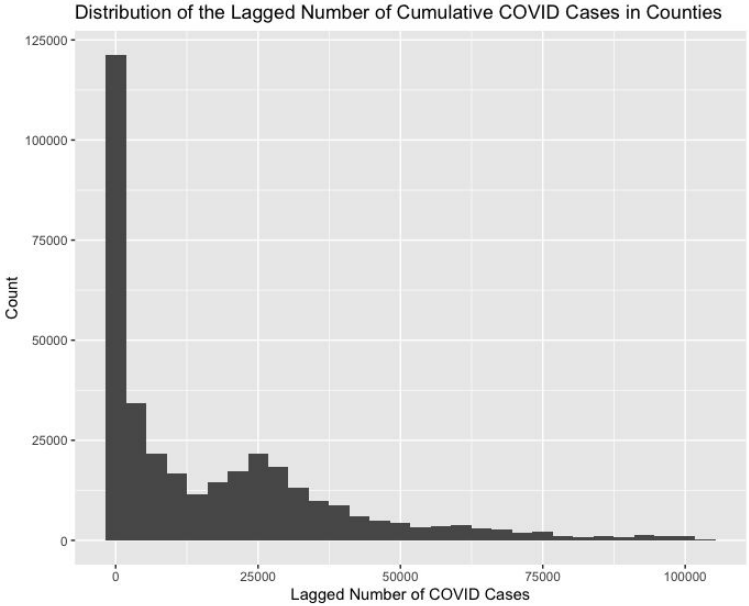
Final Dataset Processing:

- *Yelp Dataset* :
  - 352,260 restaurant reviews of 26,259 businesses
  - an average of 13.4 reviews per business (the minimum number of reviews is 1, and the maximum is 626).
- *CDC and New York Times Dataset* :
  - Cumulative county-level daily COVID case counts with our dataset on businesses and reviews using the county identifiers and dates of the reviews
- *Exploratory and Distribution* : (See next Page)
- *Last Step:* Combine two datasets together.

DATA — Distribution Overview



- Reviews appear to be biased in a positive direction as there are far more 5-star reviews than any other number of stars
- People appear to post more when they've had either an extremely positive or extremely negative experience.



- The average number of lagged cumulative cases is 17,384, and the median is 8,742.
- Vast majority of cumulative caseloads are fairly low.

METHODOLOGY — (1)

- Text processing:
  - Removed non-alphabetic characters and excess whitespace, lowercased all the text and removed common English language stopwords(Toolki), stem the words in each review(Porter Stemmer), generated a dictionary of terms used in the reviews and transformed the reviews into word vectors (or a bag of words)
- Sentiment Analysis :
  - **VADER** (Hutto & Gilbert, 2014) and **TextBlob**
  - VADER uses a collection of words coded for sentiment direction and intensity by MTurk workers. Additionally, it recognizes a variety of sentence features which can affect the direction and intensity of the sentiment conveyed by a word such as negating words or all-caps lettering. It uses these rules and lexicon to determine the sentiment conveyed by a text.
  - TextBlob relies on WordNets and part-of-speech tagging to match words in texts to words with scored sentiments. These sentiments are then averaged within documents to generate an overall sentiment rating of the text.

METHODOLOGY — (2)

- Latent Dirichlet Allocation:

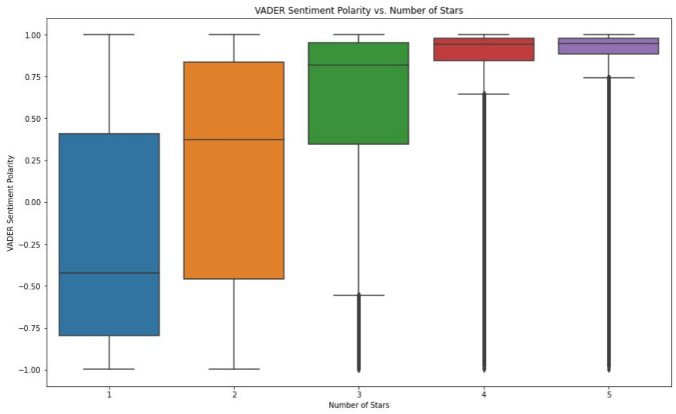
  The first step of this analysis was to determine how many topics we should use. We automated this step of the process by implementing LDA several times using a range of values for the number of topics. For each implementation of LDA, we calculated the coherence score of the model to find the number of topics which yielded the highest coherence score. Coherence scores essentially capture the extent to which words within a topic are used in similar contexts (Newman et al., 2010). After finding the highest coherence score, we performed LDA using this number of topics and examined the results. We also classified each text under the various LDA-generated topics, placing each review into the highest-probability topic from the model.

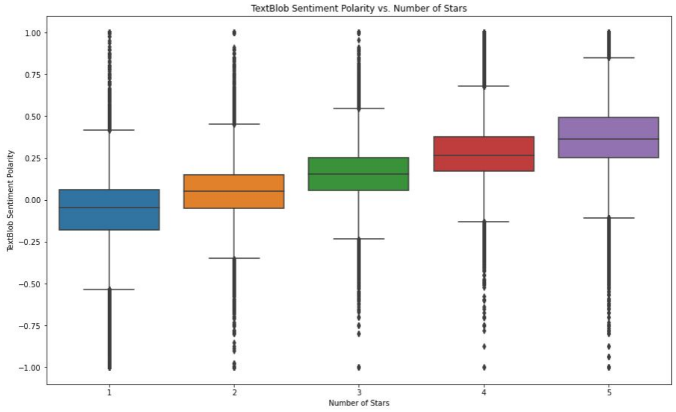
- Regression Analysis:

  Multilevel modeling allows us to properly account for the hierarchical structure of the data by fitting varying intercepts (random effects) for the different levels and also, pools some information across levels (Gelman & Hill, 2007). Using this technology, we are able to estimate the effects of county-level COVID case counts on the sentiment of reviews within businesses.

RESULTS — Model Checking

- Sentiment Analysis **— VADER and TextBlob**
  - Plotted against numeric “star” ratings, there is a monotonic increase in sentiment scores of texts, demonstrating that the sentiment analysis is capturing the intended concepts.
  - Reviews with more stars tend to score higher on sentiment polarity while reviews with fewer stars tend to score lower on sentiment polarity.
  - People are not only posting positive reviews at higher rates but also appear to be using somewhat positive language across the spectrum of their ratings.

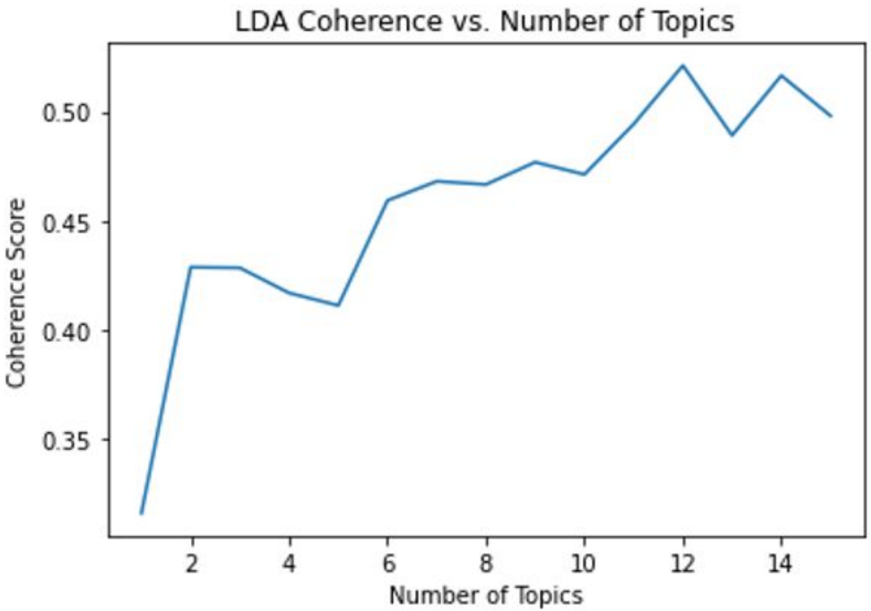




RESULTS — Model Checking

- **Latent Dirichlet Allocation — Coherence Score**
  - The aim of LDA is to capture real topics in the text that make sense to human interpreters. One such measure of whether the topics make sense is coherence. As noted in the methodology section, we fit several LDA models using a range of the number of topics from 1 to 15.
  - We can see the coherence of each LDA model plotted against the number of topics.
  - This plot indicates that coherEULTSpeaks around 12 topics.

    Thus, in our final results we use 12 topics in our LDA model.

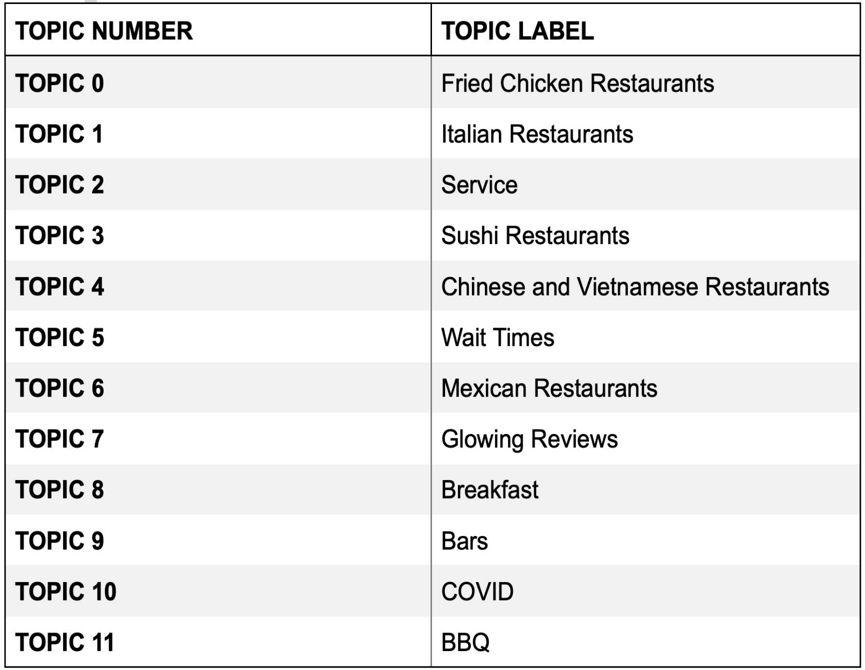


RESULTS — Model Checking

- **Latent Dirichlet Allocation — Word Topics**
  - we display word clouds of the top 10 words for each topic—most of which seem to fit together logically
  - Ex: Topic 1 “crust,” “pizza,” and “pasta.”

    Topic 2 descriptions of service at the restaurants.

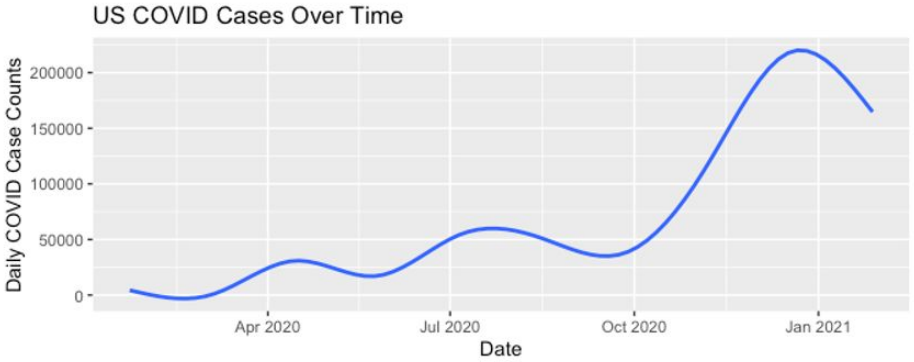
    Topic 10 “covid,” “mask,” “distance,” and “outside.”
  - In the right tae below, we label the 12 topics with descriptive words that match the words in the word clouds.



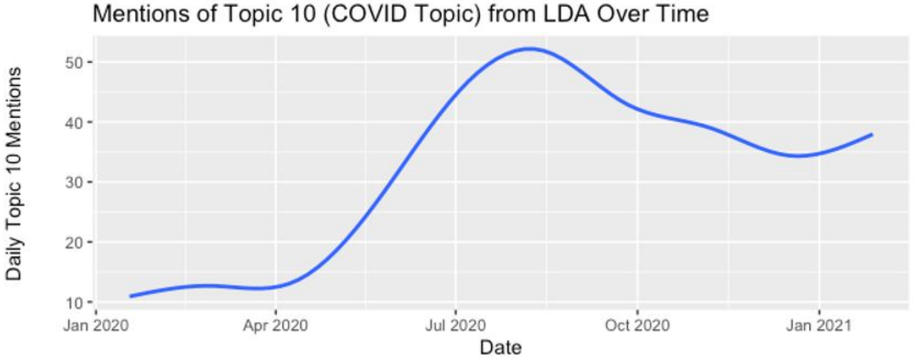
TOPIC NUMBER	TOPIC LABEL
TOPIC 0	Fried Chicken Restaurants
TOPIC 1	Italian Restaurants
TOPIC 2	Service
TOPIC 3	Sushi Restaurants
TOPIC 4	Chinese and Vietnamese Restaurants
TOPIC 5	Wait Times
TOPIC 6	Mexican Restaurants
TOPIC 7	Glowing Reviews
TOPIC 8	Breakfast
TOPIC 9	Bars
TOPIC 10	COVID
TOPIC 11	BBQ

RESULTS — Main Analysis

- **Number of Reviews Mentioning COVID and COVID Cases**
  - In the plot , we can see that COVID mentions do seem to mirror the COVID cases—with something of a lag.(Ex, during the second wave over the summer of 2020, we see the largest peak of COVID mentions in Yelp reviews.)
  - Additionally, following the third wave of COVID cases in December 2020, we see the beginning of an upswing in COVID mentions in January 2021.
  - It appears, therefore, that people are concerned about COVID and that this concern spills over into how they interact with restaurants. Importantly, concern seems to track the risk of COVID measured using daily case counts.



US COVID Cases Over Time



Mentions of Topic 10 (COVID Topic) from LDA Over Time

RESULTS — Main Analysis

- **COVID and Review Sentiment**
  - As we can see in Table 2, the lagged number of COVID cases appears to significantly drive VADER sentiment polarity in a negative direction.
  - Interestingly, the lagged COVID cases appear to drive TextBlob sentiment polarity significantly in a positive direction.
  - TextBlob does have a slightly lower correlation with star ratings than VADER sentiments (0.69 vs. 0.65), so this might be an artifact of some errors in TextBlob.

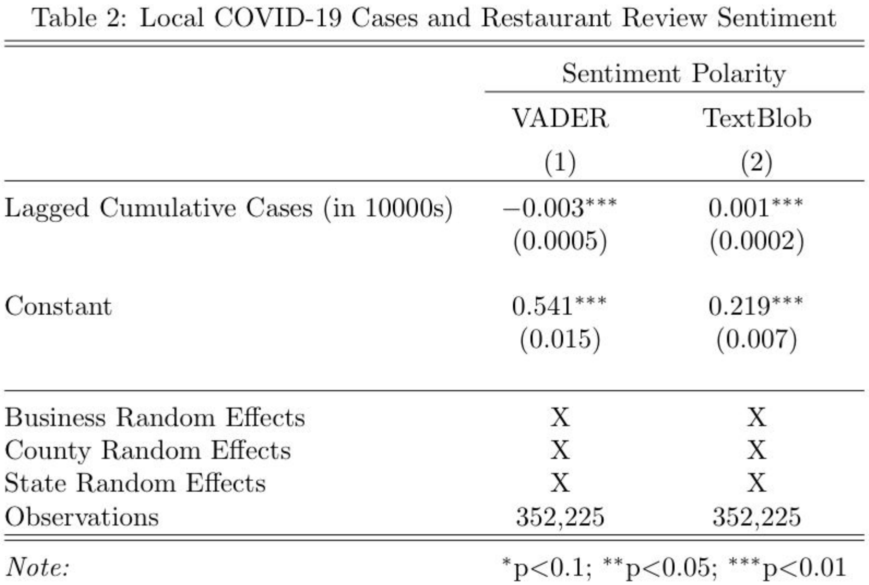


Table 2: Local COVID-19 Cases and Restaurant Review Sentiment

	Sentiment Polarity	
	VADER	TextBlob
	(1)	(2)
Lagged Cumulative Cases (in 10000s)	-0.003^{***}	0.001^{***}
	(0.0005)	(0.0002)
Constant	0.541^{***}	0.219^{***}
	(0.015)	(0.007)
Business Random Effects	X	X
County Random Effects	X	X
State Random Effects	X	X
Observations	352,225	352,225

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

RESULTS — Main Analysis

- **COVID and Review Sentiment – Regression Analysis Setup**
  - **Question:** :

    How restaurant owners might reduce some of the negative reviews from COVID.?

  - **Hypothesis** :

    Features of restaurants that are more COVID-safe might reduce the impact of COVID on restaurant ratings and the sentiments expressed in restaurant reviews.

  - **Regression Modeling** :
    - we fit multilevel linear models, predicting both VADER and TextBlob sentiment with varying intercepts (random effects) for counties, states, and businesses.
    - Additionally, we included the lagged number of cumulative cases in each county for each day in 10,000s, the restaurant features mentioned above, and an interaction between lagged COVID cases and restaurant features.

Table 3: Local COVID-19 Cases, Restaurant COVID Accommodations, and Restaurant Review Sentiment

	Dependent variable:			
	VADER Sentiment Polarity			
	(1)	(2)	(3)	(4)
Lagged Cumulative Cases (in 10000s)	−0.003*** (0.001)	−0.005*** (0.001)	−0.002*** (0.001)	−0.00004 (0.001)
Outdoor Seating	0.074*** (0.007)			
Lagged Cumulative Cases * Outdoor Seating	0.001 (0.001)			
Takeout		−0.095*** (0.014)		
Lagged Cumulative Cases * Takeout		0.002 (0.001)		
Delivery			−0.146*** (0.012)	
Lagged Cumulative Cases * Delivery			−0.0003 (0.001)	
DriveThru				−0.350*** (0.021)
Lagged Cumulative Cases * DriveThru				−0.011*** (0.003)
Constant	0.510*** (0.012)	0.629*** (0.019)	0.632*** (0.015)	0.514*** (0.037)
Business Random Effects	X	X	X	X
County Random Effects	X	X	X	X
State Random Effects	X	X	X	X
Observations	327,127	344,001	342,469	50,165
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 4: Local COVID-19 Cases, Restaurant COVID Accommodations, and Restaurant Review Sentiment

	Dependent variable:			
	TextBlob Sentiment Polarity			
	(1)	(2)	(3)	(4)
Lagged Cumulative Cases (in 10000s)	0.001*** (0.0003)	0.001 (0.001)	0.001*** (0.0004)	0.002*** (0.001)
Outdoor Seating	0.032*** (0.003)			
Lagged Cumulative Cases * Outdoor Seating	−0.0002 (0.0004)			
Takeout		−0.032*** (0.006)		
Lagged Cumulative Cases * Takeout		0.0004 (0.001)		
Delivery			−0.059*** (0.005)	
Lagged Cumulative Cases * Delivery			−0.0005 (0.0005)	
DriveThru				−0.138*** (0.007)
Lagged Cumulative Cases * DriveThru				−0.006*** (0.001)
Constant	0.205*** (0.006)	0.248*** (0.009)	0.257*** (0.007)	0.206*** (0.014)
Business Random Effects	X	X	X	X
County Random Effects	X	X	X	X
State Random Effects	X	X	X	X
Observations	327,127	344,001	342,469	50,165
Note:		*p<0.1; **p<0.05; ***p<0.01		

RESULTS — Main Analysis

- **COVID and Review Sentiment – Regression Analysis Main Findings**
  - **Findings:**
    - In Table 3, we find that only one of the interactions is significant, and it is significant in a negative direction, implying that having a Drive-Thru actually makes the impact of COVID worse for restaurants. However, we also see that the interactions with outdoor dining and takeout are positive, in the expected direction, meaning that they work to reduce the negative impact of COVID on review sentiments.
    - We find similar results in Table 4, with only the Drive-Thru interaction significant—but again, significant and negative.
    - In both Tables 3 and 4, the coefficient estimates for takeout, delivery, and Drive-Thru options are negative and significant, meaning that restaurants with takeout, delivery, and Drive-Thru tend to receive more negative reviews using both TextBlob and VADER sentiment polarity.

RESULTS — Main Analysis

- **COVID and Review Sentiment – Regression Analysis Main Findings**
  - **Findings:**
    - Stepping back to examine what these results tell us, both tables imply that restaurants should, to the extent possible, implement an outdoor dining option as this can increase the sentiment polarity in reviews.
    - Restaurants with outdoor dining have reviews with more positive language than those that do not have this option.
    - Additionally, the results using VADER sentiment polarity suggest that outdoor dining options can reduce the negative impact of COVID cases in the county on reviews sentiment.
    - Moreover, restaurants should not adopt takeout, delivery, or Drive-Thru options because these features are associated with more negative sentiments in reviews.

REFERENCE

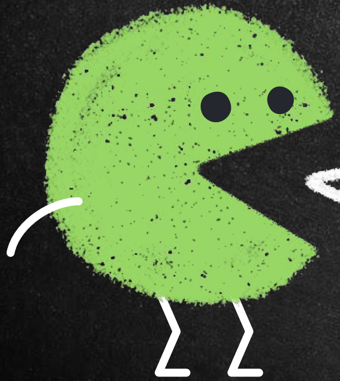
- Centers for Disease Control and Prevention. (2021). Trends in Number of COVID-19 Cases and Deaths in the US Reported to CDC, by State/Territory. Retrieved December 15, 2021, from https://covid.cdc.gov/covid-data-tracker/#trends_dailycases
- Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI, June 2014.
- National Restaurant Association. (2021). Restaurant Industry Facts at a Glance. Retrieved December 17, 2021, from https://restaurant.org/research-and-media/research/industry-statistics/national-statistics/.
- Newman, D., Lau, J.H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 100–108. Los Angeles, California.
- Petersen, Victoria. (2021). The Cold Is Coming. Will New Yorkers Still Eat Outdoors? *The New York Times*. Retrieved from https://www.nytimes.com/2021/12/02/dining/new-york-winter-outdoor-dining.html.
- Stabley, Justin (2021). How Restaurants Have Innovated to Face the Pandemic. *PBS NewsHour*. Retrieved from https://www.pbs.org/newshour/economy/how-restaurants-have-innovated-to-face-the-pandemic.
- The New York Times. (2021). Coronavirus (Covid-19) Data in the United States. Retrieved December 15, 2021, from https://github.com/nytimes/covid-19-data.
- Yelp. (2021). Yelp Open Dataset. Retrieved December 14, 2021, from https://www.yelp.com/dataset.

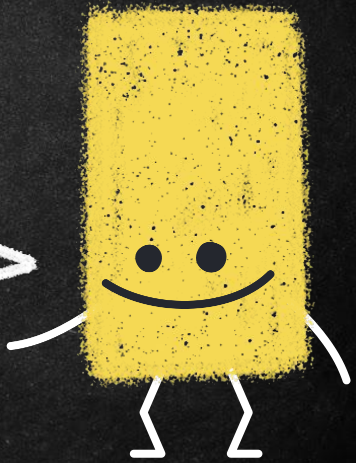
Appendix-Version 1

GR5067

NATURAL LANGUAGE PROCESSING SOCIAL SCIENCES

Final Project Presentation

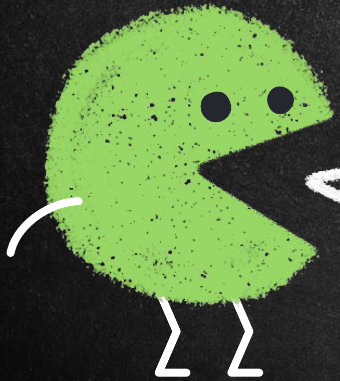




Group 15:

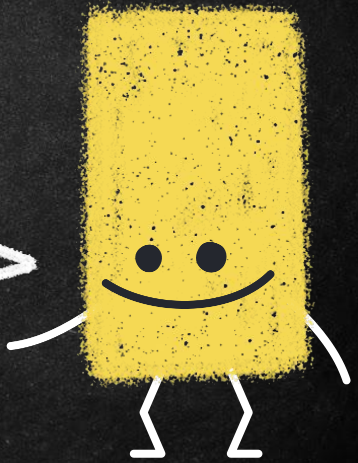
Yanji DU Junjie Ma Yixuan Li

Samuel Frederick Xiaojia Liu



TOPIC:

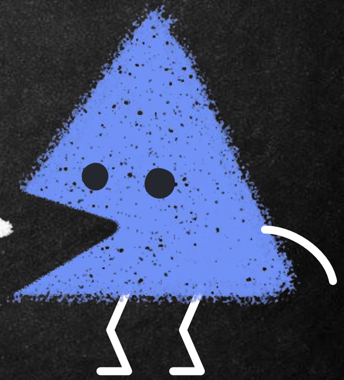
NLP FOR YELP REVIEWS DATA

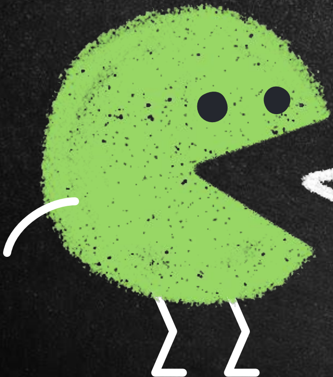


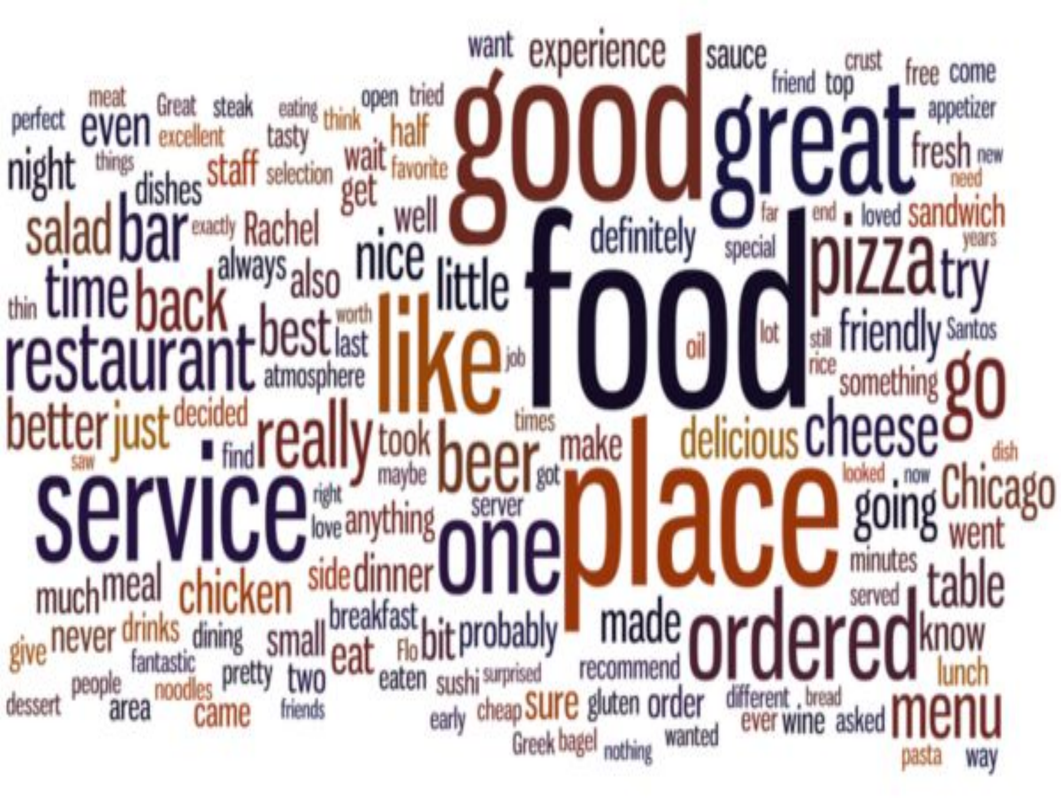
Overview: Background

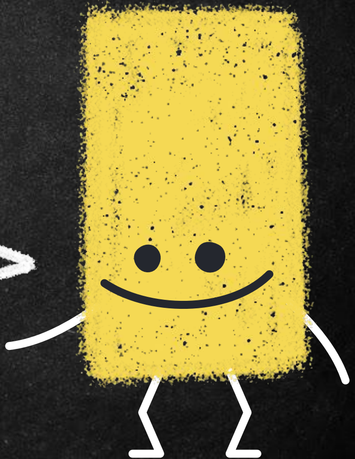
- ➜ Yelp is an application to provide the platform for customers to write reviews and provide a star rating. Reviews on Yelp.com can be an important factor in driving customers to a business.

- ➜ We hope our final project can turn this unstructured data into useful insights, which can help restaurants better understand how customers like or dislike their food or services






want
experience
sauce
crust
free
come
friend
top
meat
Great
steak
eating
open
tried
perfect
even
excellent
tasty
think
half
good
great
appetizer
fresh
new
need
night
things
dishes
staff
selection
wait
favorite
get
salad
bar
exactly
Rachel
well
definitely
far
end
loved
sandwich
years
special
nice
time
always
also
little
food
pizza
try
thin
back
worth
friendly
Santos
restaurant
best
last
like
job
oil
lot
still
rice
atmosphere
something
go
better
just
decided
times
delicious
cheese
saw
find
really
took
make
dish
maybe
beer
got
looked
now
service
right
server
place
Chicago
love
anything
one
going
went
minutes
much
meal
chicken
side
dinner
table
served
breakfast
made
ordered
know
give
never
drinks
dining
small
eat
Flo
bit
probably
fantastic
pretty
two
eaten
surprised
recommend
lunch
dessert
people
noodles
sushi
sure
gluten
order
different
bread
area
came
friends
early
cheap
ever
wine
asked
menu
Greek
bagel
nothing
wanted
pasta
way



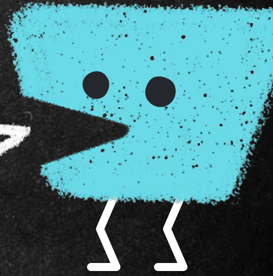
Overview: Data

The dataset we choose comes from a kaggle competition directly held by Yelp.

We use two major datasets, which are the *yelp_academic_dataset_business* and *yelp_academic_dataset_reviews*

Important Features:

- Business_id
- Review_star
- Text
- Date
- Name
- Stars
- Review_count



Data

March 2021

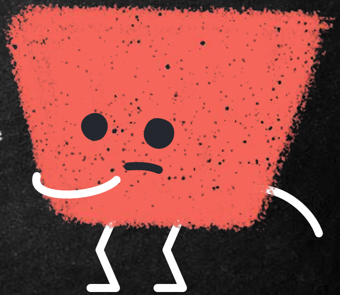
Most recent dataset for 2020-2021

267145 Yelp Reviews

Large sample size across 8 metropolitan areas

2 Crucial features

Good to enhance the accuracy of the model

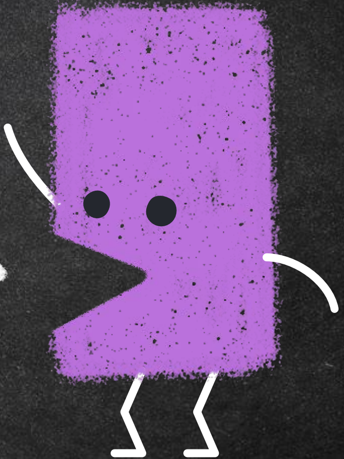


“

Analysis Methods:

1. Exploratory Data Analysis
2. Text pre-processing
3. Sentiment analysis
4. TF-IDF Analysis

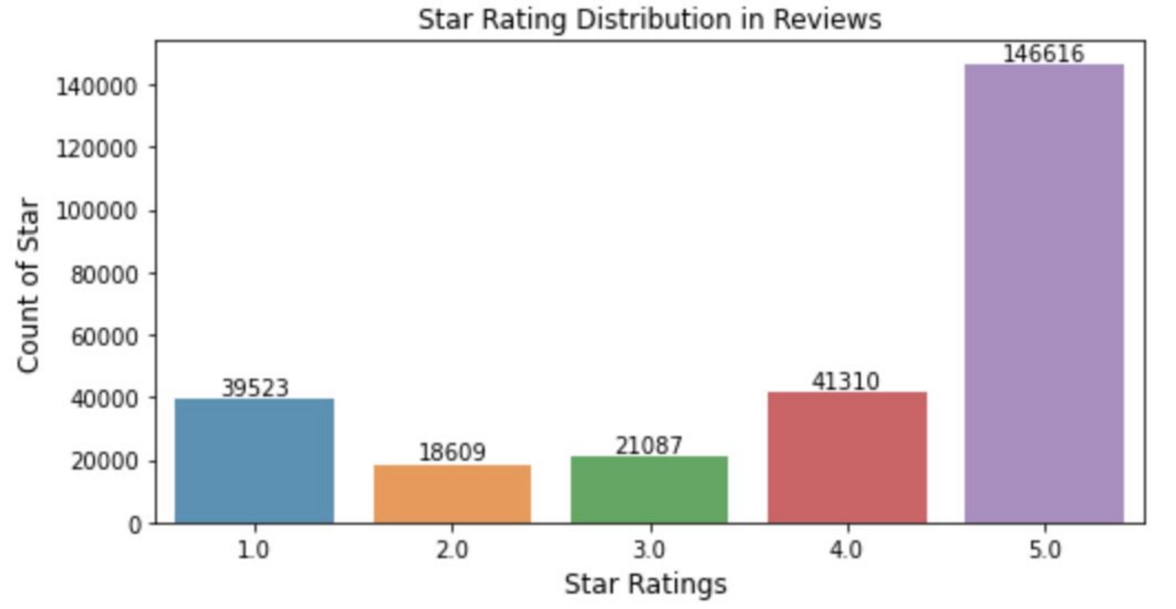
(term frequency–inverse document frequency)

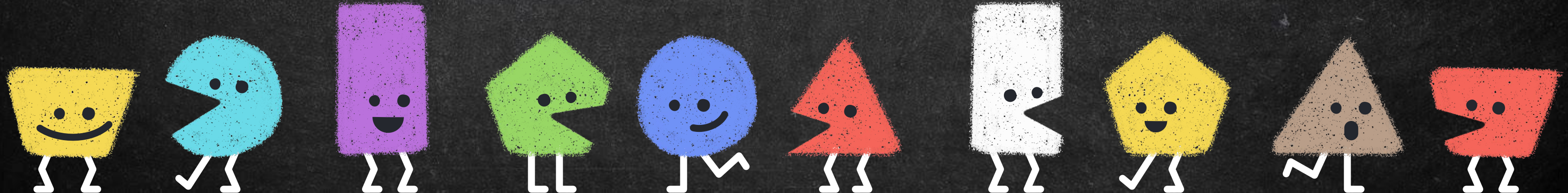


Analysis Methods – EDA

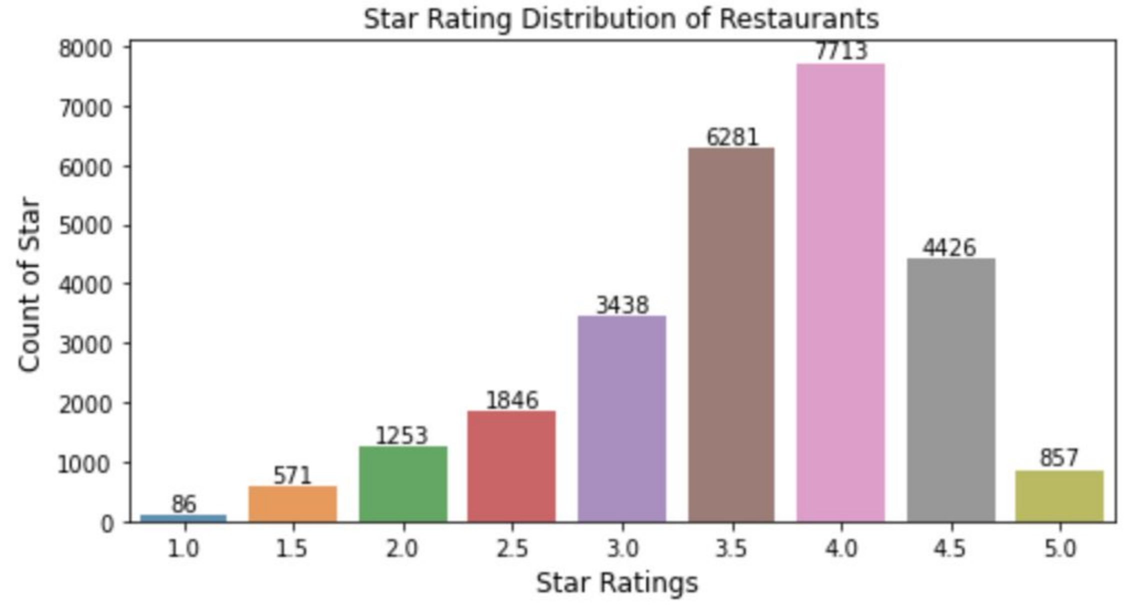
- Most popular is 5-star reviews

- 1-star reviews are more common than 2- or 3-star reviews.

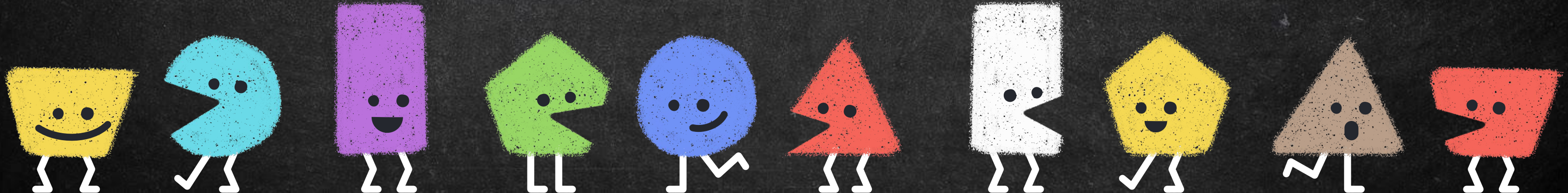




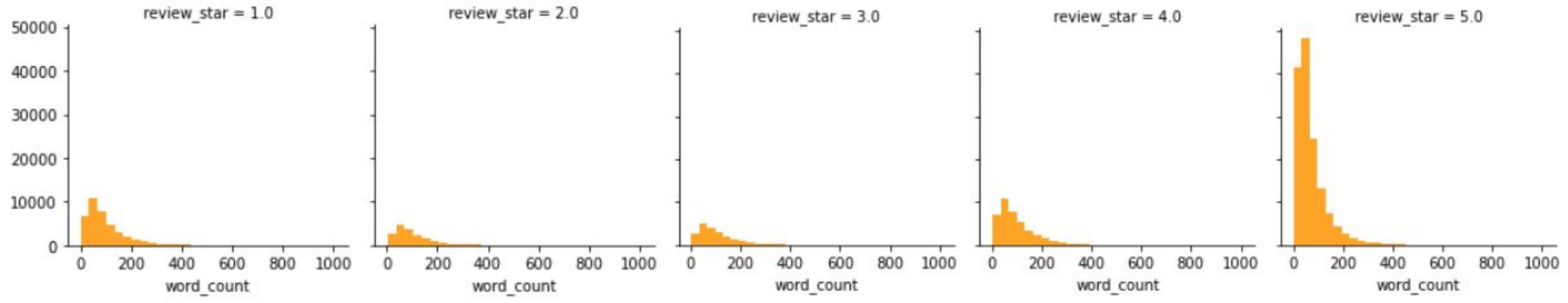
Analysis Methods – EDA



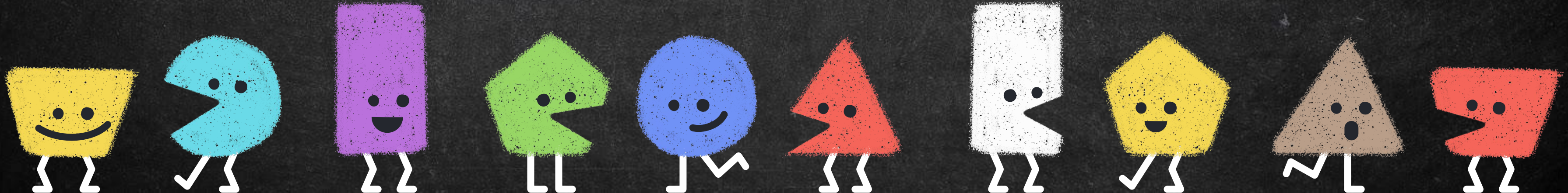
- Most of the restaurants' stars are around 3.5 to 4.



Analysis Methods - EDA



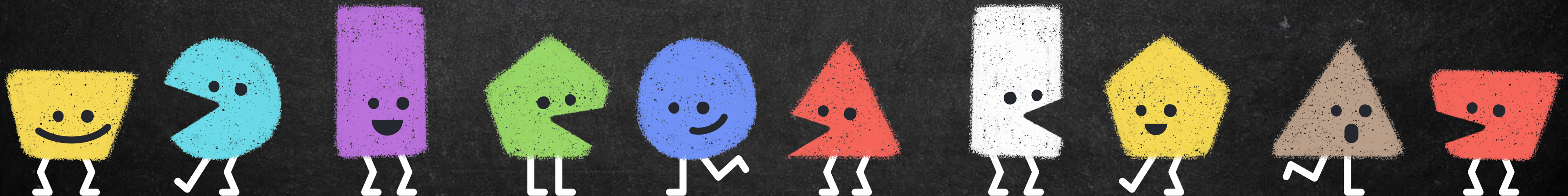
- Skewed to the right
- Most of the reviews are within 100 words
- Barely no reviews exceed 400 words



Analysis Methods – Text Pre-processing

Clean the text data:

- Lowercase all words, remove punctuation, stopwords and numbers
- Remove non-English reviews
- Remove other stopwords (“get”, “us”, “im”, “ive”……)
- Lemmatize

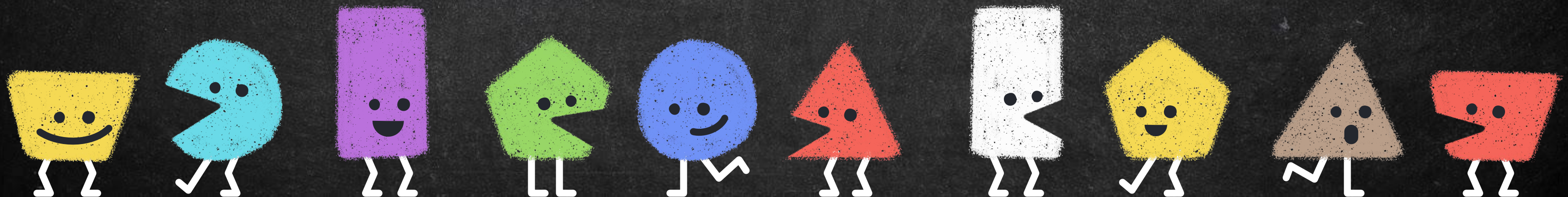


Analysis Methods - Text Pre-processing

Create Dependent Variable:

- 4-star reviews and 5-star reviews are positive sentiment
  1-star reviews and 2-star reviews are negative sentiment

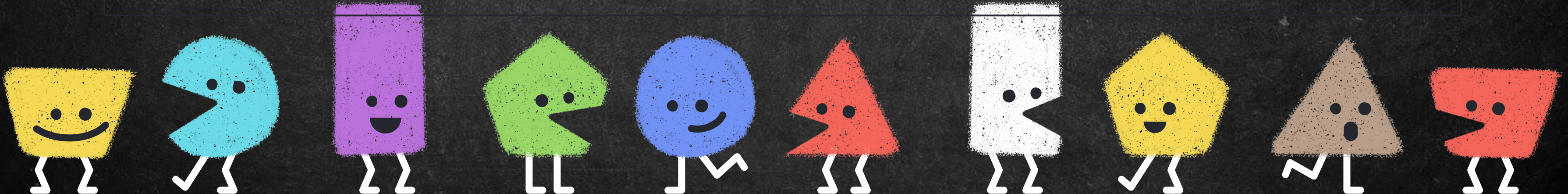
- Including 3-star review might affect accuracy



Analysis Methods – Text Pre-processing

Create Input Variables (Bag of Words):

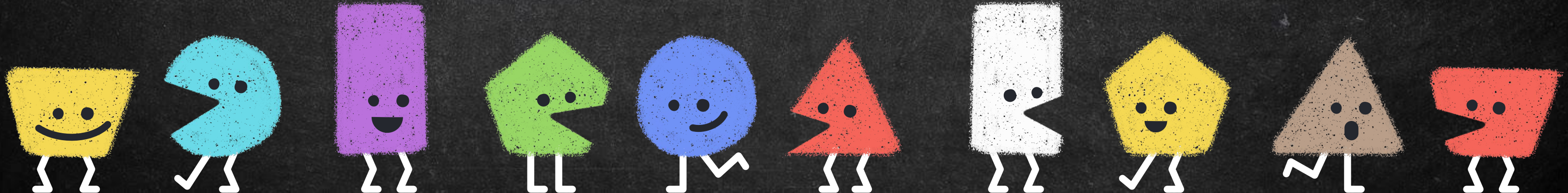
- Frequency of words in corpus
- Vector representation of words into number 0 and 1
- but it does not contain information on important words.



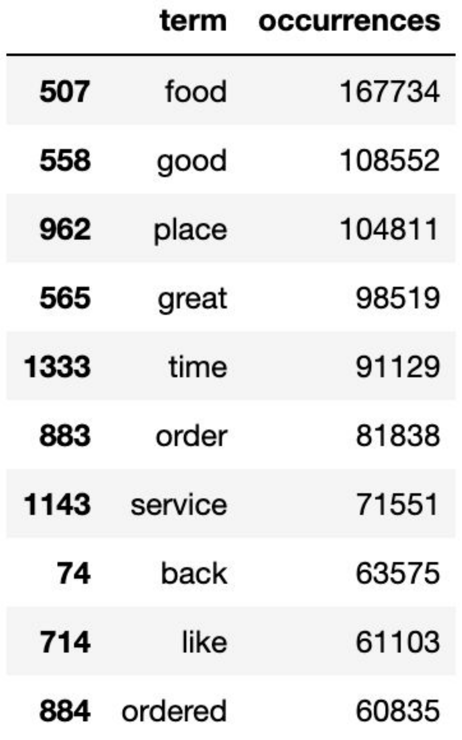
Analysis Methods - Text Pre-processing

Calculating TF-IDF:

- a scoring measure widely used in information retrieval (IR) or summarization.
- TF-IDF is intended to give weight to reflect how important a term is in a given document



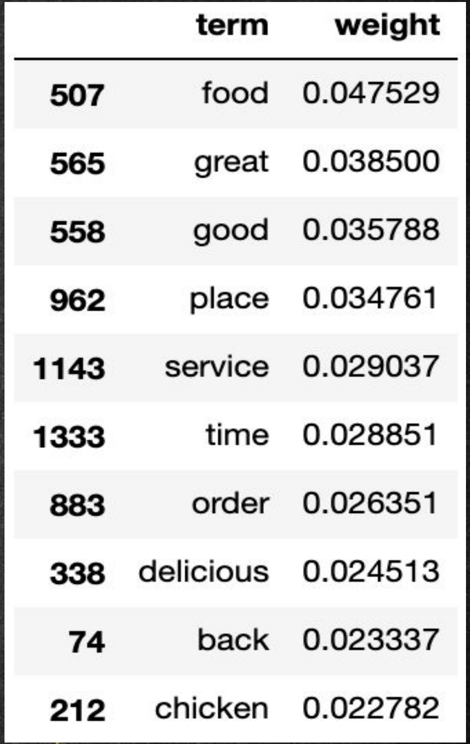
Bag of Words:



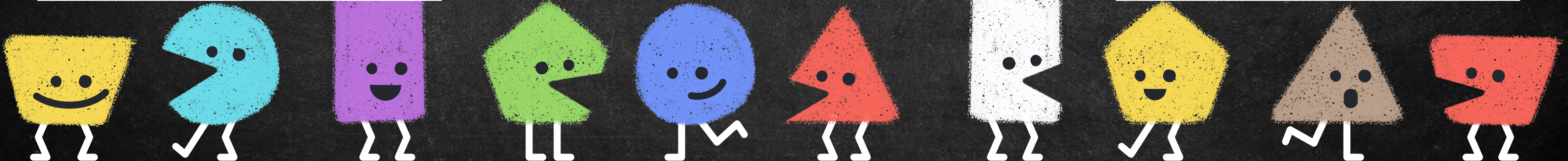
	term	occurrences
507	food	167734
558	good	108552
962	place	104811
565	great	98519
1333	time	91129
883	order	81838
1143	service	71551
74	back	63575
714	like	61103
884	ordered	60835

- Word “delicious” is very crucial though the frequency is not in top 10
- Word “like” is not given that much importance though its frequency is high

TF-IDF weight:

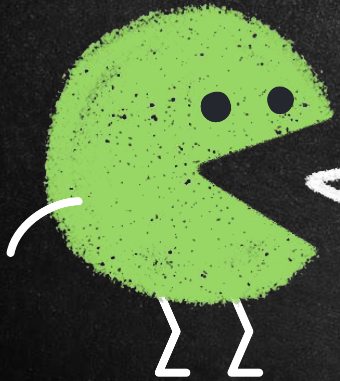


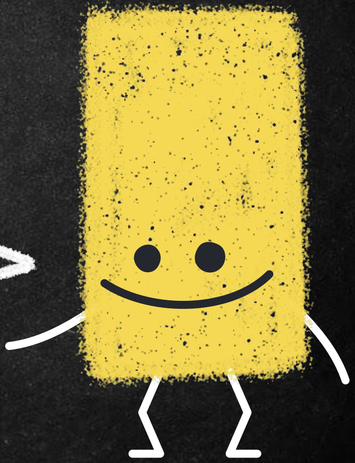
	term	weight
507	food	0.047529
565	great	0.038500
558	good	0.035788
962	place	0.034761
1143	service	0.029037
1333	time	0.028851
883	order	0.026351
338	delicious	0.024513
74	back	0.023337
212	chicken	0.022782



Sentiment analysis

1. Classification Tree
2. Model Interpretation
3. Model Enhancement - Cross Validation
4. TF-IDF Model

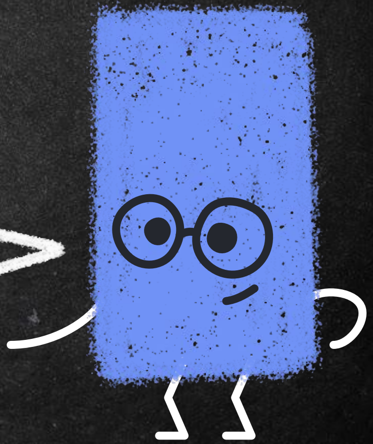




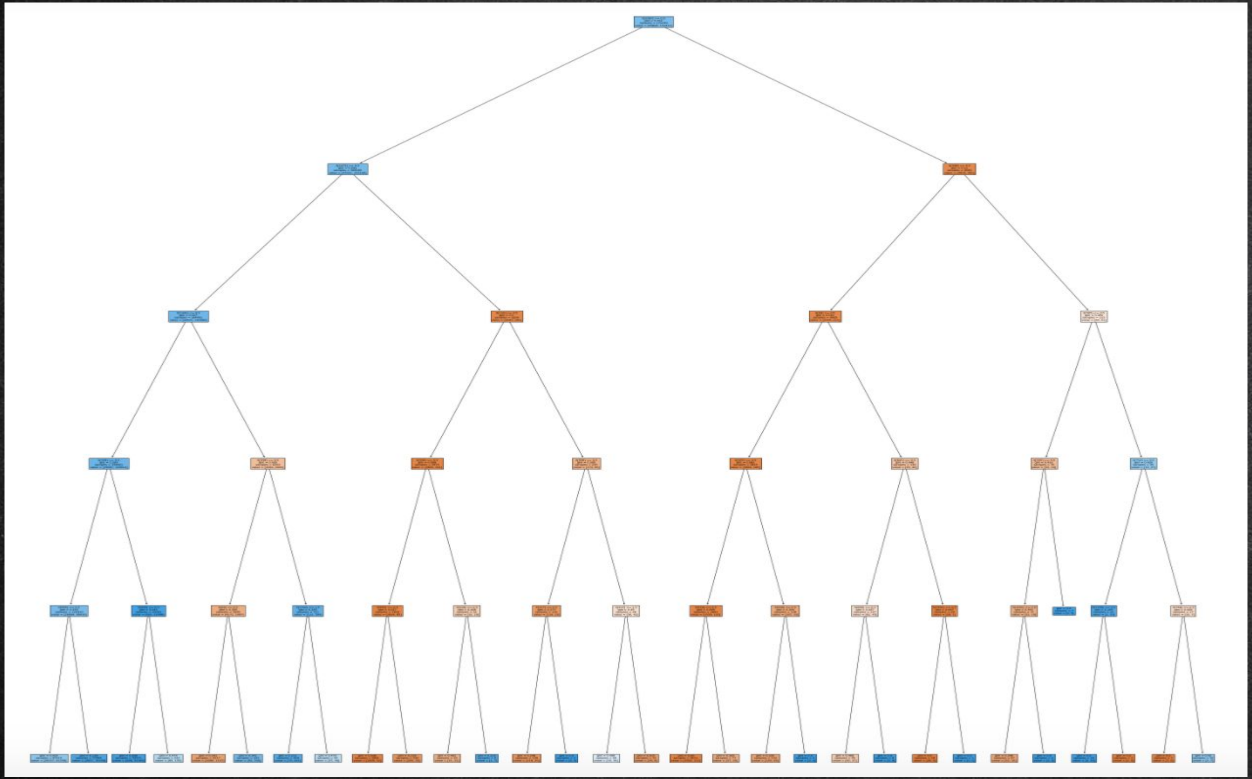
Sentiment Analysis -- Classification Tree

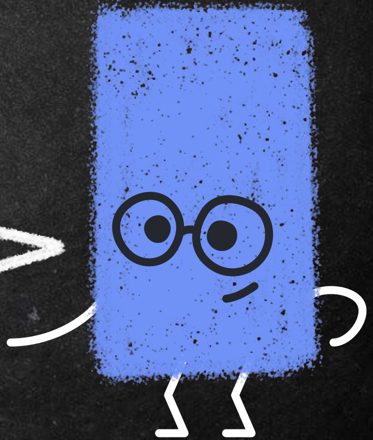
- First we split the data into a training (70%) and testing (30%) dataset.

- Building CART Model :
  CART Model accuracy: 0.8174564469370614



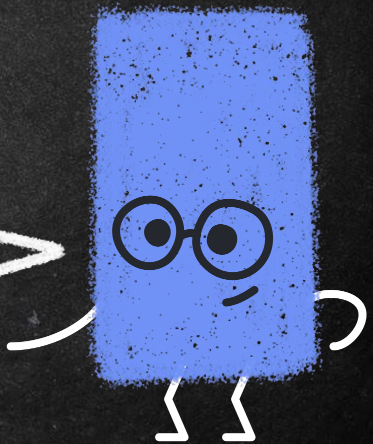
Classification Tree





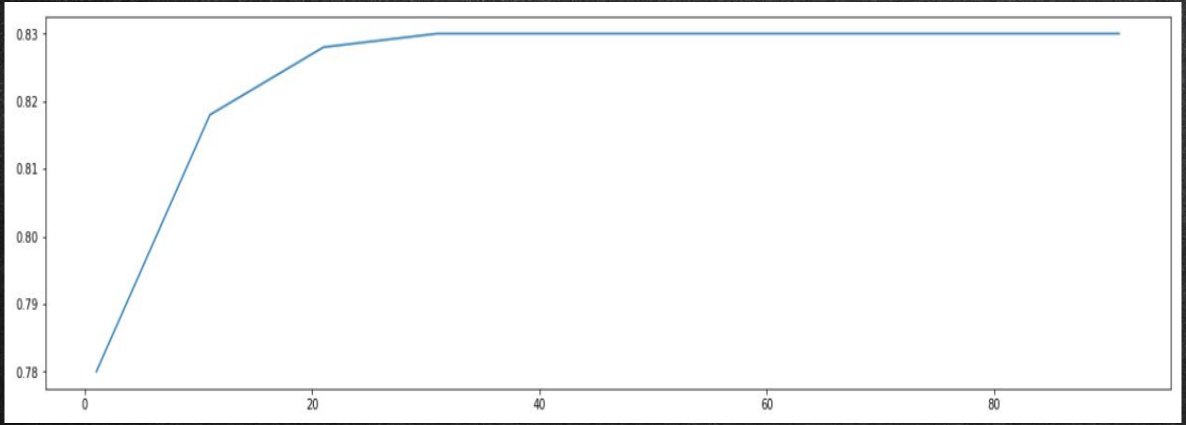
Sentiment Analysis
-- Cross Validation

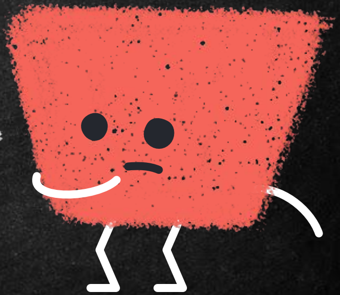
- Altering the default number of splits
- Accuracy for improved model: 0.84280256847923
  84.28% test-set accuracy



TF-IDF Model

- Accuracy for improved model: 0.889742339266845 88.97% test-set accuracy





Discussion

1. Current: high accuracy of our model
   - Reviews data not like text from Twitter.
   - Reviews data might be ideal for sentiment
   - Using rating as a predictor

2. Future: could do the analysis for different restaurants
   - In order to reveal the key aspects of that restaurant that drive overall customer perception
   - Could examine variation by locality

Conclusion

1. We used **exploratory data analysis** to assess the public perception of restaurants on Yelp and found that **5-star reviews are the most common reviews**. Most restaurants had stars around 3.5-4.5.
2. We used **Bag of Words** and **TF-IDF** to interpret the frequency of the specific word and weigh the importance of each one to better understand. We got the 18792 positive value counts so the **dataset is biased towards positive reviews**.
3. We built machine learning model which accurately predicts the **sentiment of reviews** with the **accuracy around 89%**. Our future goal is to use our trained sentiment model to predict different dataset.

Citation:

1. Yelp, I. (2021, March 2). Yelp dataset. Kaggle. Retrieved December 7, 2021, from https://www.kaggle.com/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_business.json.

1. Luca, M. (n.d.). Reviews, reputation, and revenue: The case of yelp. Retrieved December 7, 2021, from https://www.hbs.edu/ris/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf.

Thanks for watching :)