

Analysis on the Survivability of NYC Restaurants During COVID-19

Date: Dec. 2020

Ling Ao, Lei Zhang, Szu-Yu Chen

Content

- Introduction
- Data collection & exploration
- Methodology & Results
 - Basic features classification
 - Word to Vector Embedding
 - Sentiment score & TextBlob
 - Negative reviews clustering / LDA topic model
- Discussion
- Conclusion
- Limitation & Recommendation



Introduction

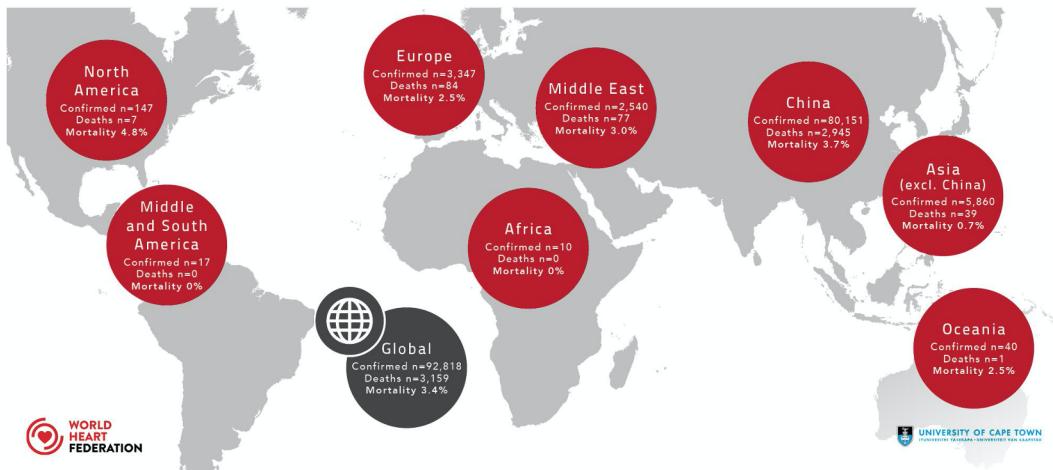
Motivation

1. The COVID-19 pandemic has a profound influence on all countries and all the business over the world.
2. Food industry is one the businesses in the U.S. that got hit by the pandemic the hardest.
3. In this paper, we are going to explore the factors that correlate with the survivability of restaurants

Coronavirus COVID-19

Global map of SARS-CoV-2/COVID-19 epidemic by region

Data extracted on 03 March 2020 at 23:00 CET from the online interactive dashboard, hosted by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, Baltimore, MD, USA (<https://github.com/CSSEGISandData/COVID-19>). Not included on the map are 706 cases tested positive for SARS-CoV-2 on "Diamond Princess" cruise ship, 6 people died (0.8%).





Introduction

Research questions:

Research questions 1: What features are most likely correlate with the survivability of restaurants located in New York City when facing pandemic like COVID-19?

Research question 2: Whether the extracted risk assessment features can be applied to restaurants located in other comparable areas and predict their survival state during 2020?





Introduction

Background: Yelp

- ❖ Crowd-sourcing websites to help customers to target local business
- ❖ Million of customers rely Yelp for food hunting
- ❖ Review data in yelp is reliable and up-to-date and has wide coverage of all businesses





Data Collection

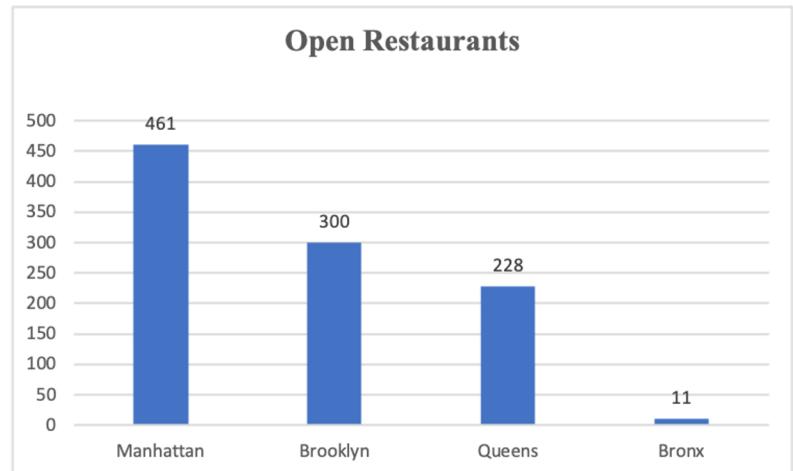
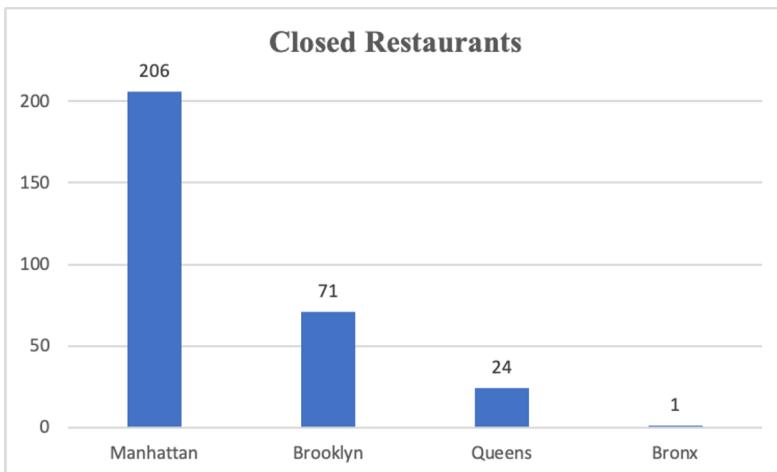


- ❖ Permanent close restaurants
due to COVID-19
- ❖ May 8th to October 31st, 2020
- ❖ Total 302 restaurants
- ❖ 1000 open restaurants
- ❖ Using Yelp API



Data Exploration

Area Distribution

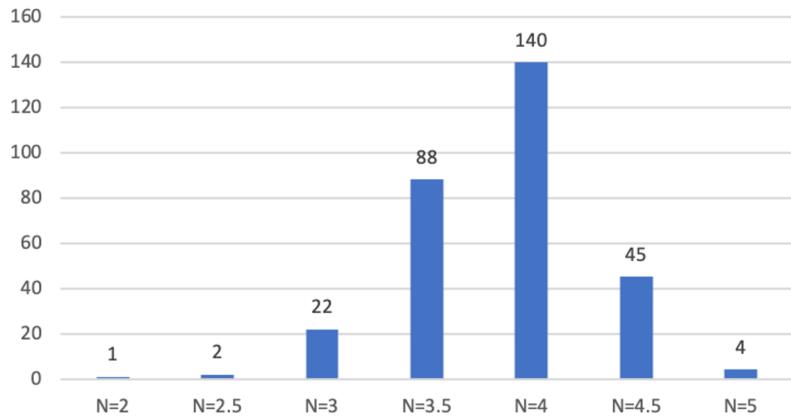




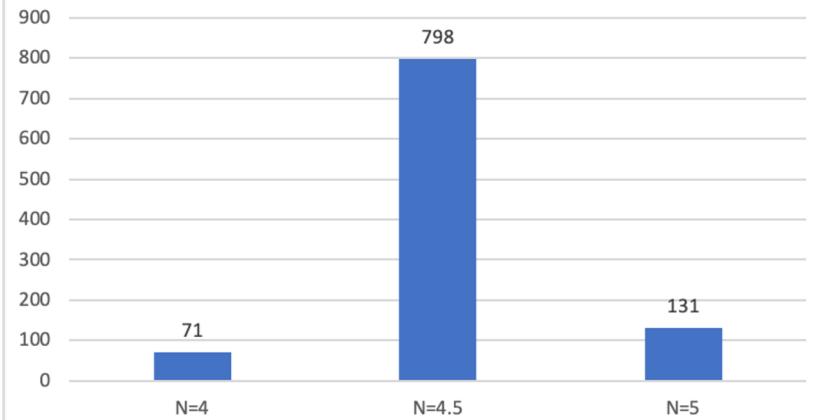
Data Exploration

Rank Distribution

Rank Distribution _ Closed



Rank Distribution _ Open





Data Exploration

Price

1 stands for \$, 2 stands for \$\$, 3 stands for \$\$\$, and 4 stands for \$\$\$\$.

Is_close

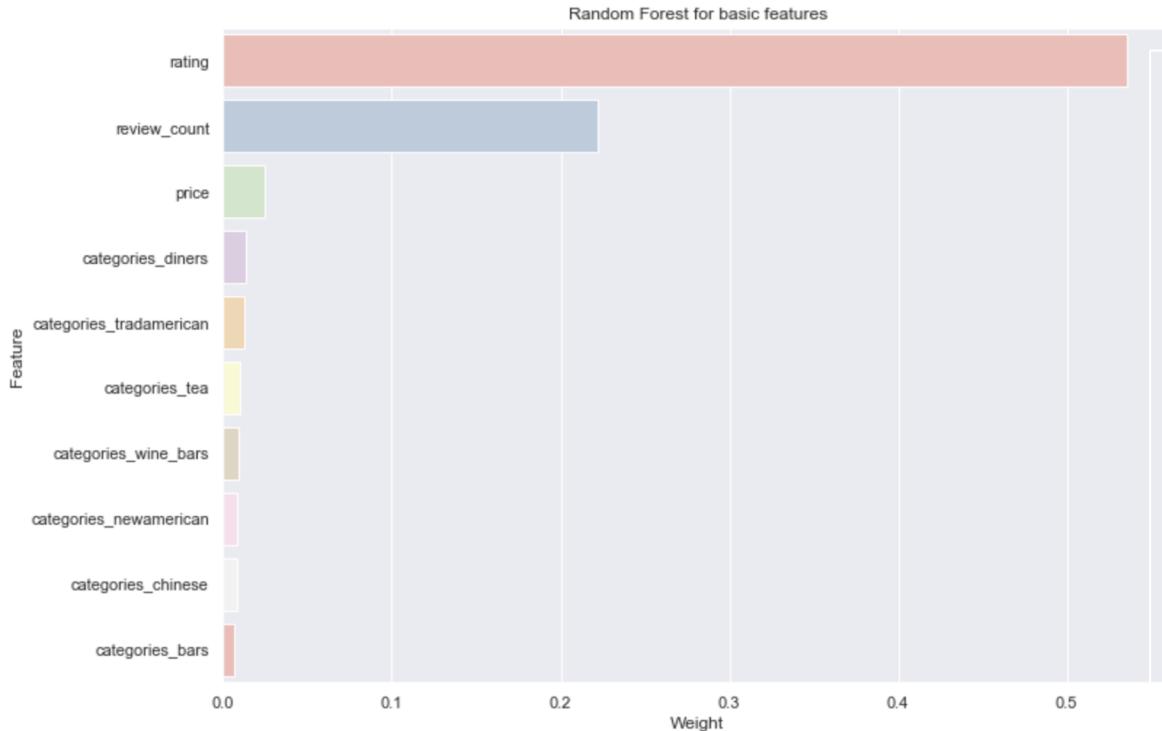
0 stands for closed and 1 stands for open

Review

10 reviews from the official website of the restaurant.



Method_Basic features classification

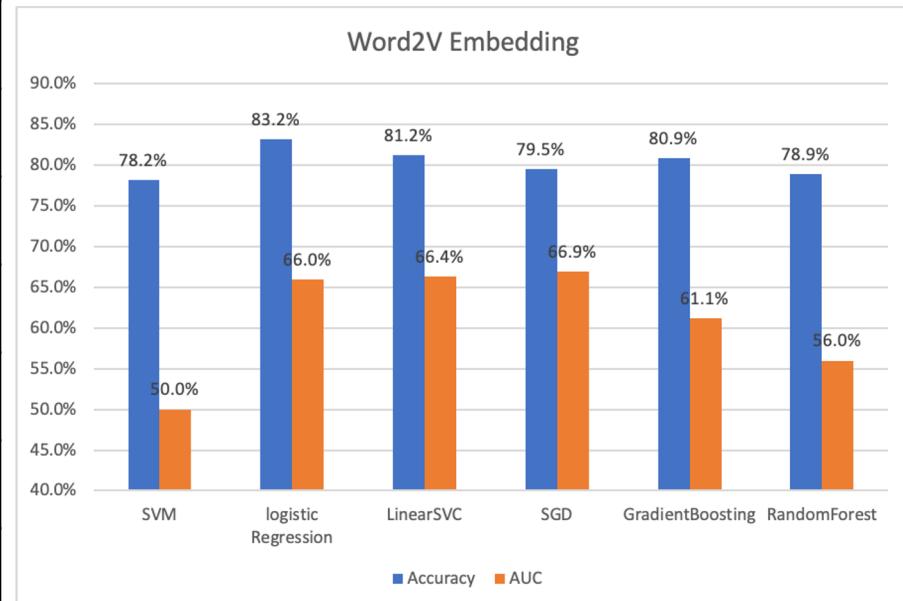


- Use original features on Yelp to build a classification model
- Use Random Forest Model to fit and find feature importance
- The model performance is good, AUC: 0.906; F1 macro: 0.8573
- Traditional American, tea, wine bars, bars, New American, and Chinese seem to be the top survival features



Method_Word to Vector Embedding

Embedding Method	Model	Accuracy	AUC
Word to Vector	SVM	78.19%	50.00%
Word to Vector	logistic Regression	83.22%	65.98%
Word to Vector	LinearSVC	81.21%	66.35%
Word to Vector	SGD	79.53%	66.94%
Word to Vector	GradientBoosting	80.87%	61.15%
Word to Vector	RandomForest	78.86%	55.98%



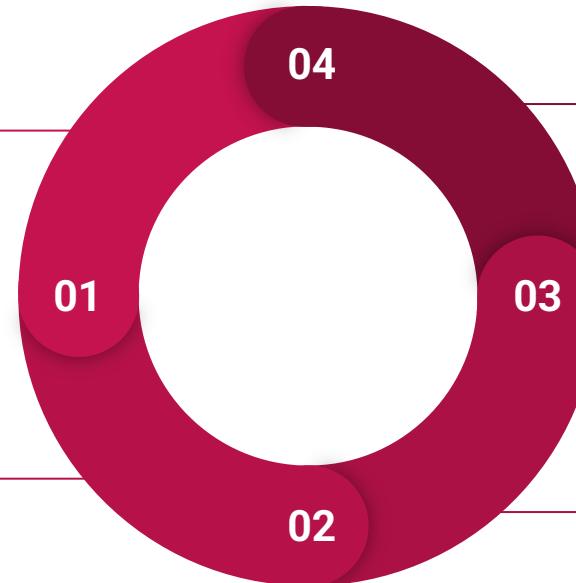


Method_Text analytics

Use tf-idf, doc2vec in five classification model

logistic regression, linear SVC, SGD, Gradient Boosting, and random forest

Linear SVC has the highest test accuracy: 0.7379



The test accuracy slightly increased to 0.7629



Second adding: sentiment scores (pos, neg, neu, com)



Method_TextBlob

- Dependent variable: is_closed
- Independent variables :
review_count, rating, price, reviews,
- Reviews: polarity and subjectivity

Model	Accuracy
RandomForest	93.18%
SVC	73.86%
LogisticRegression	84.85%
LinearSVC	77.27%
SGD	73.48%
GradientBoosting	93.56%



Method_Negative reviews clustering

- Use nltk vader to polarize if compound score < 0 then negative
- Nearly 10% of reviews are negative, we use reviews of closed restaurants only for clustering

Cluster 1

20 minutes **understaffed**
hasty delayed finally order
misunderstood good lot
finally swapped food
delayed food excellent
hasty order lot
miscommunication
server misunderstood food
rushing understaffed lot
rushing service **ignored**
said definitely

Cluster 2

food isn 30 minutes **left**
hungry menu price lunch
menu noodles known modify
website gotten noodles
noodles saw **terrible** fresh
known **modify receipt**
charged paid noticed regular
price ordered noodles noodles
terrible website picked
noticed receipt website
noodles price 17

Cluster 3

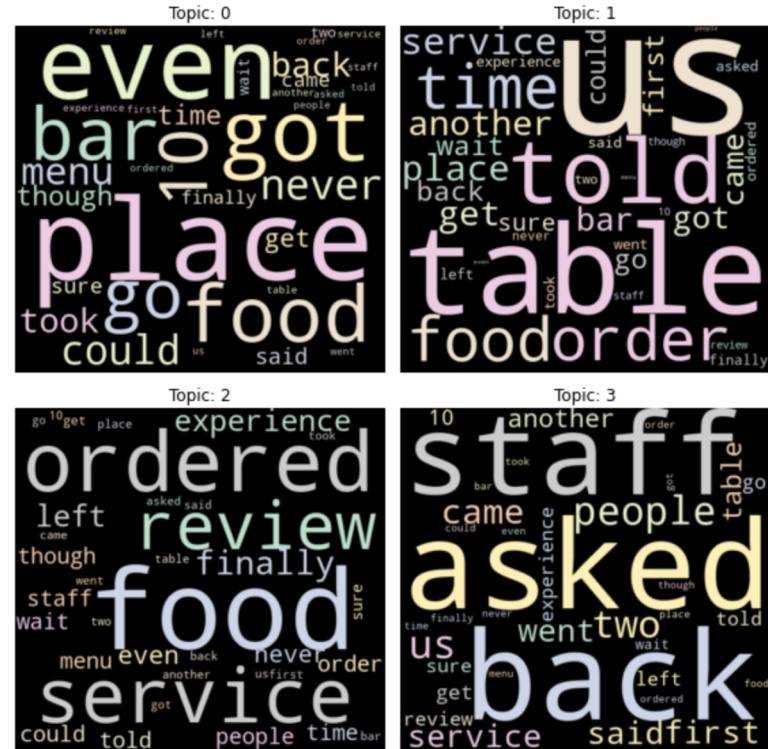
sat bar probably
microwaved food
lukewarm servers said
ignore took problem food right
problem short ribs nearby got
bother send **didn bother**
proceeded **ignore** place
quiet night place green curry
sitting nearby lukewarm didn
got special said hello
asparagus cooked

Cluster 4

mozzarella balls know place
ricotta jalapeño mediterranean
restaurant **assumes**
anorexic toddler left
choose eat **smallest**
portions payed 10 left **nut**
ordered obviously size portions
imaginable size toddler
selfish places places
ordered place owner anorexic
choose

Method_LDA topic model

- Use words at least showing five times and remove stop words and non-negative words
 - The result shows that the generated topics are regarding place, table, food, and staff





Discussion Algorithms

- **Basic feature model**
 - Random Forest: **90.64% AUC**
- **Textblob approach:**
 - GradientBoosting model: **Accuracy: 93.56%**,
- **Word to Vector embedding:**
 - SGD model: **AUC 66.94%; Accuracy: 79.53%**
 - Logistic regression : **Accuracy 83.22%; AUC:65.98%**
- **Sentiment score approach:**
 - LinearSVC model: **Accuracy 76.29%.**



Discussion_Findings

Based on the data of NYC restaurants:

- Categories belong to traditional American, bars, and tea are more likely affected by unpredictable changes
- Bars are more obvious because our model extracts the important keywords like bar and bartender

Textual data findings:

- Building effective communication is important
- Restaurants communicate about the COVID-19 crisis can create clarity, build resilience, and catalyze positive change during uncertainty

Our model predicts the Brooklyn area with AUC performance of 0.6477, decreased by the original of 0.7611. Currently, our indicators can continue to be optimized and refined. (research question 2)



Conclusion

- ❖ 2 Research questions
- ❖ 302 closed restaurants, 1000 open restaurants are used, and 10 reviews of each restaurant
- ❖ 4 different approaches
- ❖ Feature Importance : rating > review count > price > category
- ❖ This research can be used to predict the survivability of restaurants in other area models, but still can be enforced.



Limitation & Recommendation

Data size

- Only scripted 302 closed restaurants and 1000 open restaurants
- Increase the number of research restaurant and script more reviews,

Number of features

- Analyzed few features: rating, price, review count, category, reviews
- Adopt mores specific features, such as, delivery service, environment rating, service rating

Models

- Only used some baseline models
- Try more advanced feature engineering models, such as: LSTMs, BERT, Fine-turning BERT,

Reference

- [1] Chen, X. et al., 2018. "Las Vegas Business Closure Prediction Model."
- [2] Michael Luca, 2016. "Reviews, Reputation, and Revenue: The Case of Yelp.com"
Harvard Business School
- [3] M. Hoffman and D. Blei., 2010. "Online Learning for Latent Dirichlet Allocation." *Neural Information Processing Systems*.
- [4] Dan Jurafsky et al., 2014. "Narrative framing of consumer sentiment in online restaurant reviews" *First Monday*[5] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
- [5] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
- [6] Giatsoglou, Maria, et al. "Sentiment analysis leveraging emotions and word embeddings." *Expert Systems with Applications* 69 (2017): 214-224.
- [7] Yu, Liang-Chih, et al. "Refining word embeddings for sentiment analysis." *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017.
- [8] Sarkar, D. (2019). *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Berkeley, CA: Apress.

EATER data source:

A Running List of NYC Restaurants That Have Permanently Closed,
October 2020

<https://ny.eater.com/2020/10/2/21459288/nyc-restaurant-closings-coronavirus-october-2020>

A Running List of NYC Restaurants That Have Permanently Closed,
September 2020

<https://ny.eater.com/2020/9/3/21408479/nyc-restaurant-closings-coronavirus-september>

A Running List of NYC Restaurants That Have Permanently Closed
During the COVID-19 Crisis

<https://ny.eater.com/2020/5/8/21248604/nyc-restaurant-closings-coronavirus>

Thank you!

