An NLP Analysis of Yelp Reviews during COVID-19

GR5067 Final Project Report

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

Abstract

The COVID-19 pandemic has upended the lives of countless people around the globe. The disruption also led to adaptation and innovation as people learned to cope with the new reality. As we face a new variant of the coronavirus, it is important to understand how people and businesses have reacted and adapted to the pandemic. In this paper, we seek to understand how customers reacted to the prevalence of COVID, and how they thought about their experiences at restaurants–both in terms of the sentiments they expressed and the topics they discussed. Additionally, we seek to answer the question: how can restaurants successfully mitigate any negative impacts of the pandemic on customer sentiment? Our results provide an answer, showing how restaurants successfully adapted to changing customer perceptions and conclude with steps restaurants should and shouldn't take as we brace for another wave of the pandemic. We apply Latent Dirichlet Allocation to discover the topics in restaurant reviews, finding a clear COVID topic, the prevalence of which appears to track the spread of COVID nationwide. Next, we show, using sentiment analysis, that restaurant reviews get more negative as the COVID caseload in a county increases. Finally, we show that restaurants which offered outdoor seating were able to counteract some of the negative effects of COVID cases on their customer ratings.

1. Introduction

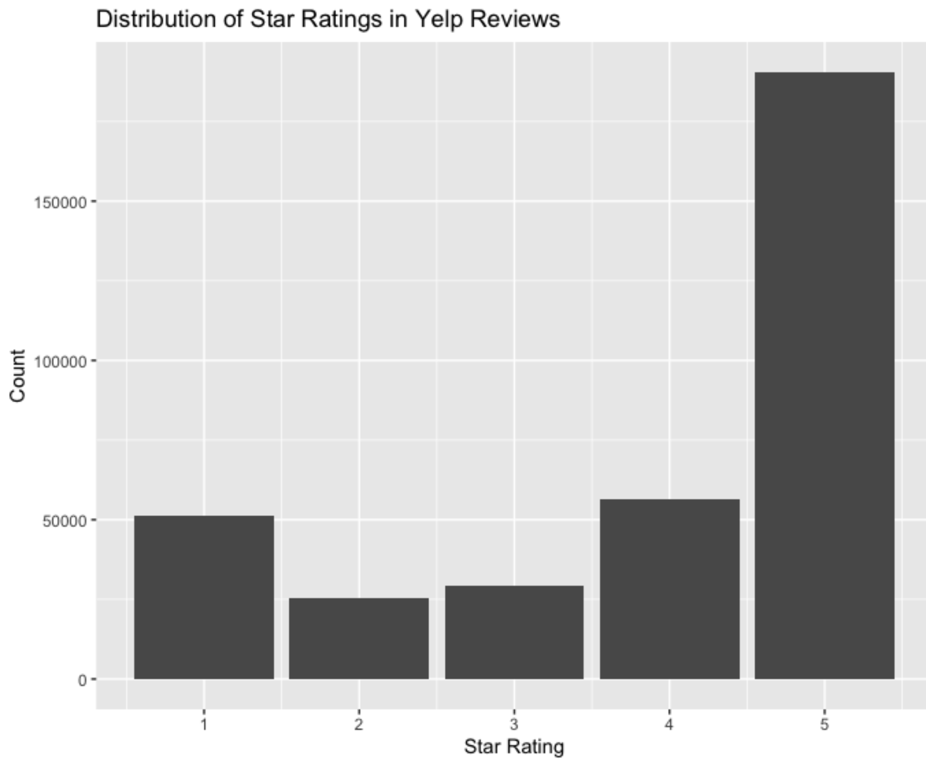
On January 18, 2020, the first case of COVID-19 was detected in the United States. Following this discovery, the virus spread quickly, killing tens of thousands of people in a matter of months. Many state governments sprung into action, issuing mask mandates and stay-at-home orders in an attempt to foster social distancing and stop the spread. Countless businesses shut down. Schools transitioned to virtual learning, and many businesses shifted to remote work. Among the hardest hit by the pandemic were restaurants. Restaurants, which typically contain many people indoors in close quarters, maskless, were unable to safely maintain social distance and were also unable to shift to remote work. Thus, restaurants were often forced to close their doors during the pandemic or to significantly reduce the number of people who could be inside. Overall, the National Restaurant Association (2021) estimates that restaurants underperformed by billions of dollars and that more than 100,000 businesses were closed at the end of 2020. Despite the negative effects of the pandemic for restaurants, some restaurants adapted, offering outdoor dining and takeout options (Petersen, 2021; Stabley 2021). As we face a new variant of the coronavirus, it is important to understand how customers thought about these changes, and further, what steps businesses might take to remain afloat as the omicron variant spreads.

We employ a dataset of customer reviews of restaurants from the website Yelp which allows people to rate restaurants and other businesses (Yelp, 2021). Using this dataset, we examined what people were discussing in their reviews of restaurants and how people were reviewing restaurants. We study whether certain features of restaurants are associated with better reviews on Yelp. Specifically, we performed an analysis using Latent Dirichlet Allocation (LDA) to recover the most coherent topics from the corpus of restaurant reviews. In fact, our LDA model returned an interpretable topic discussing COVID. Interestingly, the number of reviews which predominantly discussed the COVID topic seemed to track the number of coronavirus cases across the country. This pattern demonstrates that customers were thinking about COVID when they were visiting restaurants, and that their concern about COVID was linked to trends in coronavirus cases across the nation. Second, we performed two 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.

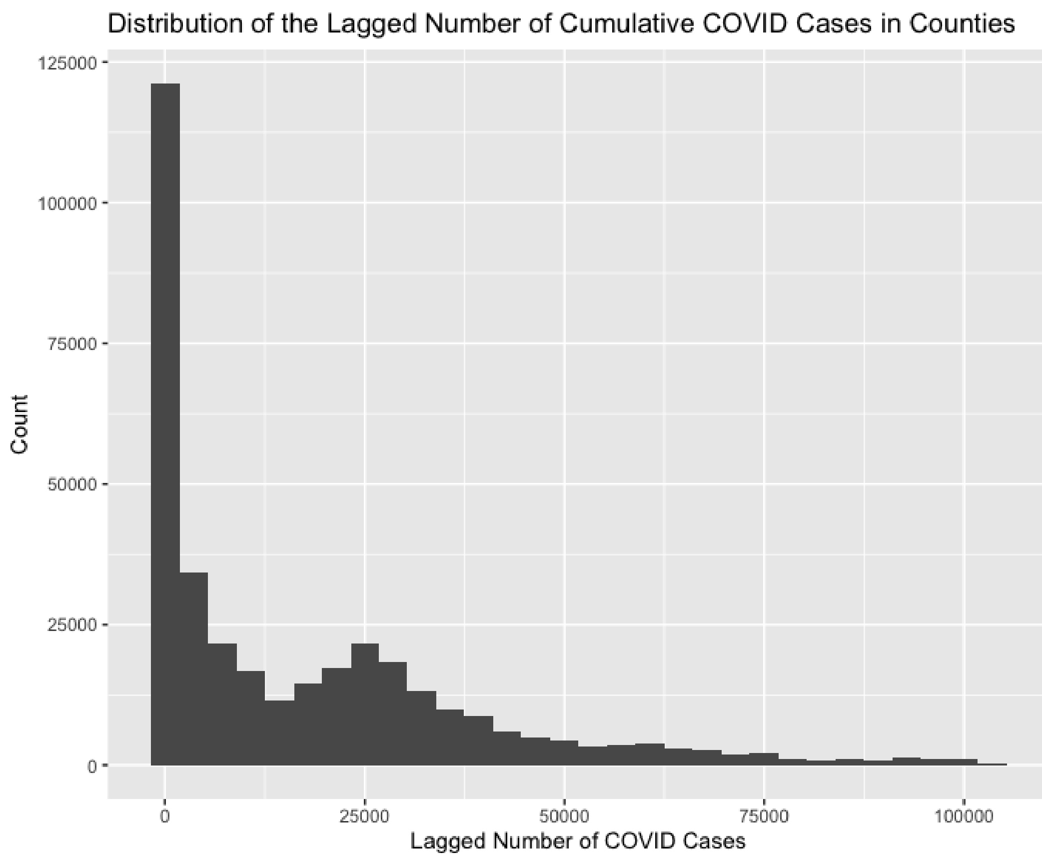
This unique dataset gives us the opportunity to investigate how COVID-19 has impacted restaurant reviews and steps restaurants can take to mitigate any negative effects of the pandemic. Our paper proceeds in four sections. We begin by laying out the various datasets employed in our analysis as well as describing our management of those data. Next, we discuss the methodology used to explore the restaurant reviews data. Then, we provide an overview of our main findings from the dataset. Finally, we conclude and offer some thoughts about the implications of our research for restaurants during the pandemic.

2. Data

In this paper, we employ data from three main sources: Yelp, the Centers for Disease Control and Prevention (CDC), and *The New York Times*. Our main data source is the academic dataset of Yelp business reviews (Yelp, 2021). There are two different datasets within the Yelp data; there is a dataset containing restaurant reviews, and there is a dataset containing information about businesses like location, hours, and whether the business offers outdoor dining. To begin, we merged the Yelp dataset containing customer reviews with the dataset containing business information, so that we could examine how attributes of the business relate to reviews data. We next removed 897 rows with no information on the businesses and removed rows which did not contain reviews of restaurants, as our project is focused on restaurant reviews. Finally, we subset the data to include only reviews that were posted on or after January 18, 2020—the date of the first confirmed COVID case in the United States–leaving us with reviews posted between January 18, 2020 and January 28, 2021. This gave us a final dataset of 352,260 restaurant reviews of 26,259 businesses with an average of 13.4 reviews per business (the minimum number of reviews is 1, and the maximum is 626). Additionally, we can see the distribution of star ratings in the reviews dataset below. This distribution indicates (1) that reviews appear to be biased in a positive direction as there are far more 5-star reviews than any other number of stars and (2) that people appear to post more when they’ve had either an extremely positive or extremely negative experience in restaurants.



The Yelp dataset of business information contained the latitudes and longitudes of each business. We used shapefiles from the US Census Bureau to locate businesses within different counties. In total, the Yelp dataset contains information on businesses from 35 U.S. counties in 11 states. The states are fairly distributed across regions with Massachusetts and New Hampshire from the Northeast, Florida, Georgia, Texas, and Kentucky from the South and Border South, Ohio from the Midwest, and Oregon, Colorado, Washington, and Wyoming from the West. Finding the specific counties allowed us to track the relationship between county-level COVID cases and the sentiment of restaurant reviews. We added the county identifiers to our dataset of businesses and reviews. Finally, we merged *The New York Times* database of cumulative county-level daily COVID case counts with our dataset on businesses and reviews using the county identifiers and dates of the reviews (The New York Times, 2021). Below, we can see the distribution of lagged cumulative COVID cases in the counties in our dataset. The average number of lagged cumulative cases is 17,384, and the median is 8,742. As we can see, the vast majority of cumulative caseloads are fairly low.



The last dataset we employ comes from the CDC. The CDC provides data on daily cases nationally. We merged this dataset with our dataset of reviews by date, so that we could compare national trends in COVID cases with national trends in restaurant reviews.

3. Methodology

In this section, we describe the approach we took to analyze the dataset described above.

3.1. Preprocessing the Texts

First, we preprocessed the reviews data from the Yelp dataset. We proceeded with our analysis in two parts: the sentiment analysis and the application of Latent Dirichlet Allocation (LDA). For the LDA analysis, we cleaned the text, removing non-alphabetic characters and excess whitespace. We lowercased all the text and removed common English language stopwords, using the Natural Language Toolkit corpus of stopwords. Finally, we used the Porter Stemmer to stem the words in each of the reviews in our dataset. Following our own cleaning of the text data, we generated a dictionary of terms used in the reviews and transformed the reviews into word vectors (or a bag of words). The dictionary was filtered to only include words used in more than 50 reviews and to exclude words used in more than 30% of the reviews. This ensured that the words included in the analysis would meaningfully distinguish among topics in the text.

3.2. Sentiment Analysis

We conducted two forms of sentiment analysis on the reviews data, using both 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. Somewhat differently, 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. The result of these two forms of sentiment analysis is a sentiment polarity score between -1 and 1 for each review in our dataset, allowing us to determine whether the sentiment in the text is positive or negative. Due to the continuous nature of the sentiment scores, we can conduct linear regressions to determine the relationships between business attributes, local COVID cases, and sentiment.

3.3. Latent Dirichlet Allocation

We used LDA to identify the major topics in our corpus of restaurant reviews in the COVID era. 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.

3.4. Regression Analyses

Because our data are hierarchically clustered, with reviews inside of businesses inside of counties inside of states, we need to account for this clustering structure to prevent the estimation of inaccurate standard errors for regression coefficients. 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.

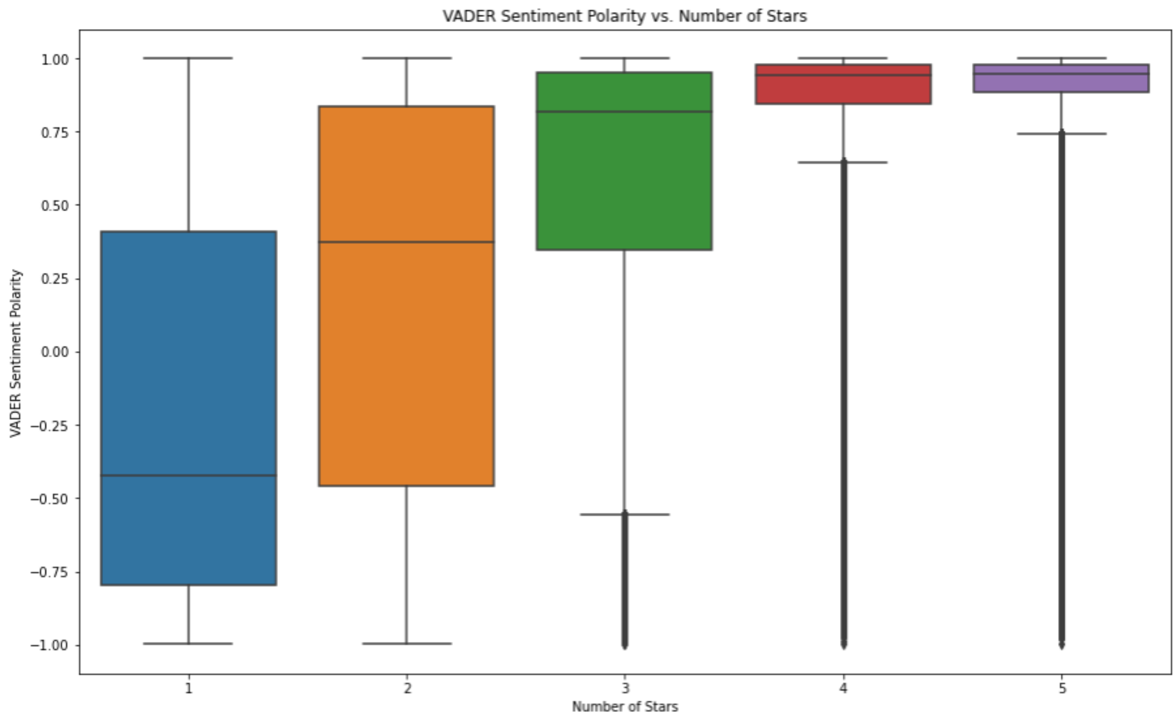
4. Results

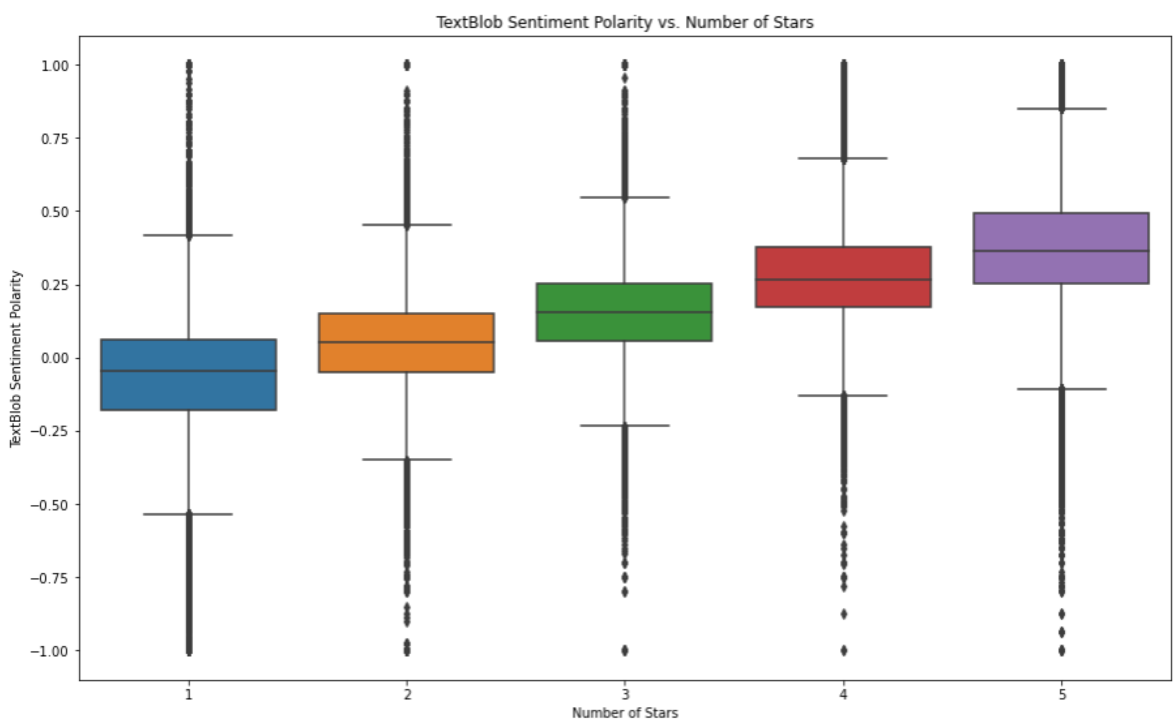
In this section, we first present model-checking results, validating the results of our sentiment analyses, examining whether the LDA topics make sense, and examining accuracy of the predictions of business characteristics. After presenting these results, we then proceed to our main analysis of business characteristics.

4.1. Model-Checking

4.1.1. Sentiment Analysis

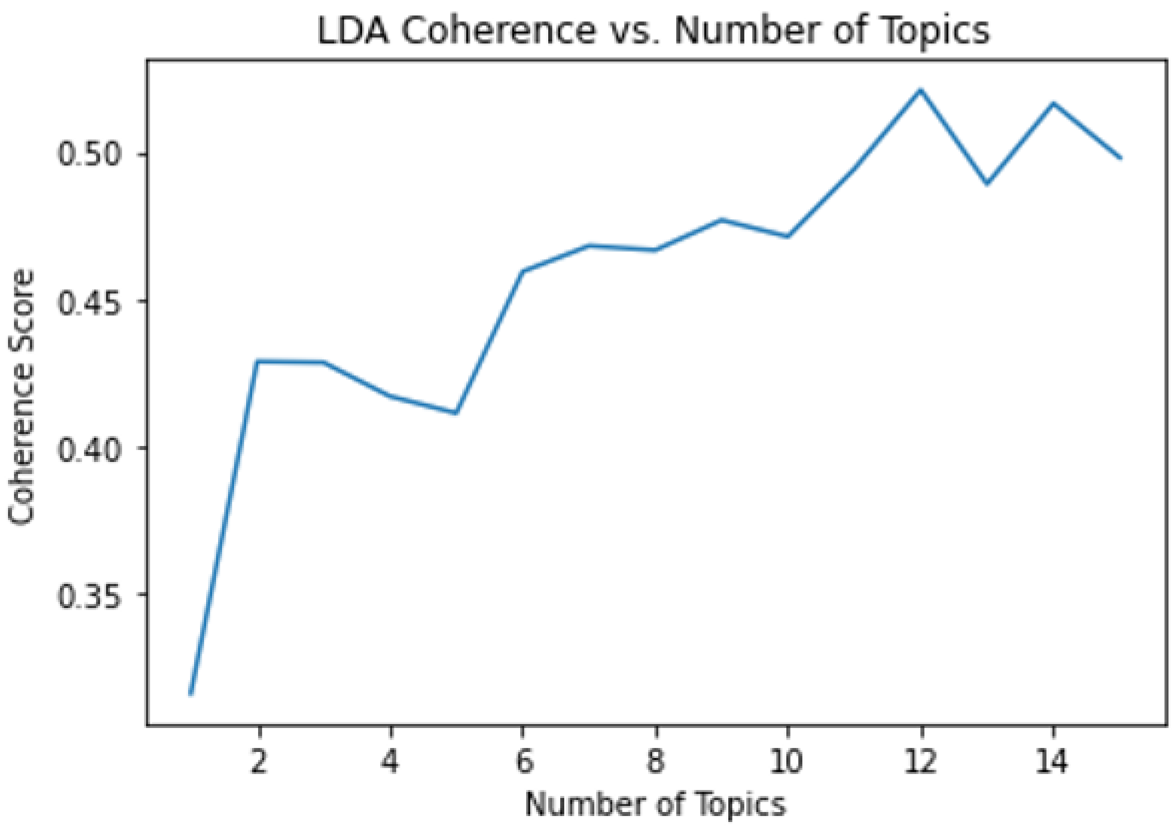
Because our data are reviews of restaurants, we have numeric ratings attached to each textual review. This allows us to compare the sentiments discerned from sentiment analyses with a numeric rating: if the sentiment analyses are accurate, reviews with more negative sentiments should also have lower numeric ratings. As we can see in the boxplots below, this is the case. 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. Admittedly, the sentiment scores do tend to be imperfect predictors of restaurant ratings; however, it also seems that reviews might be biased toward the positive end of the distribution. 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.





4.1.2. Latent Dirichlet Allocation

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. Below, we can see the coherence of each LDA model plotted against the number of topics. This plot indicates that coherence peaks around 12 topics. Thus, in our final results we use 12 topics in our LDA model.



Not only does our LDA implementation with 12 topics obtain the best coherence measure statistically, but the words in the topics returned are also interpretable. Below, we display word clouds of the top 10 words for each topic—most of which seem to fit together logically. For example, Topic 1 seems to correspond to Italian restaurants with words like “crust,” “pizza,” and “pasta.” Topic 2 seems to mostly involve descriptions of service at the restaurants. Topic 6 seems to refer mostly to food found at Mexican restaurants like tacos, burritos, and chips. Interestingly, Topic 10 mostly revolves around words related to COVID, including “covid,” “mask,” “distanc,” and “outsid.” In the table below, we label the 12 topics with descriptive words that match the words in the word clouds.

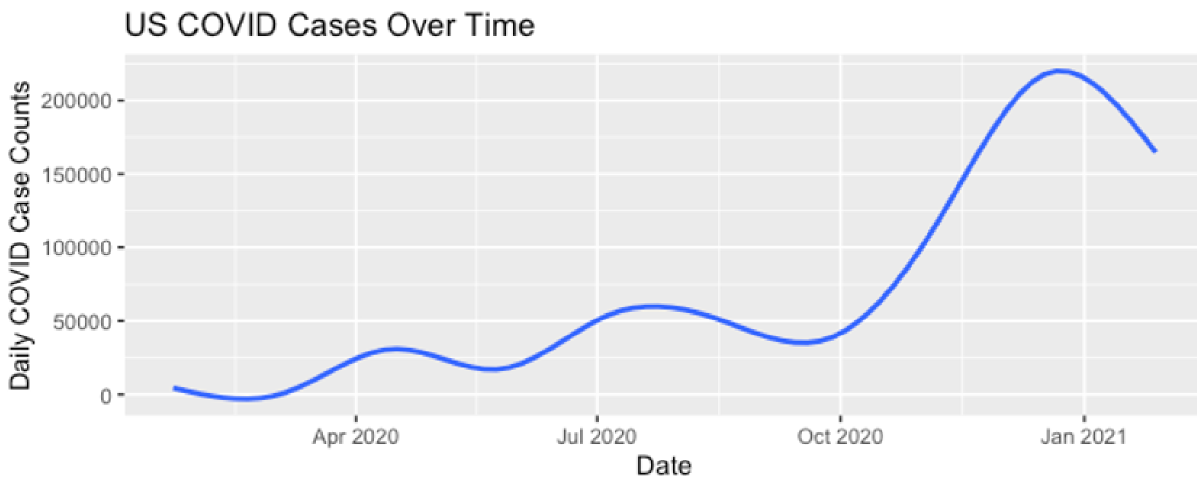
TOPIC NUMBER	**TOPIC LABEL**
TOPIC 0	Fried Chicken Restaurants
TOPIC 1	Italian Restaurants
TOPIC 2	Service
TOPIC 3	Sushi Restaurants
TOPIC 4	Chinese and Vietnamese Restaurants
TOPIC 5	Wait Times
TOPIC 6	Mexican Restaurants
TOPIC 7	Glowing Reviews
TOPIC 8	Breakfast
TOPIC 9	Bars
TOPIC 10	COVID
TOPIC 11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

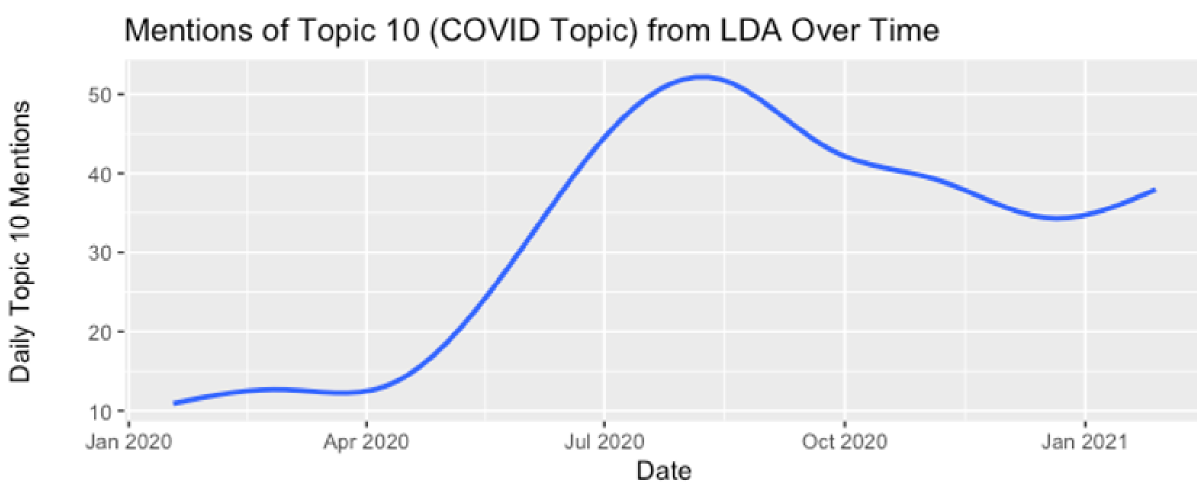

Topic 0
got cook salad sauc chicken like flavor wing fri tast
Topic 1
crust chees pasta like slice bread italian pie pizza top
Topic 2
time us servic wait came drink tabl ask back server
Topic 3
sushi tri roll like realli flavor portion dish tast price
Topic 4
chicken beef soup pho rice pork chines dumpl noodl hot
Topic 5
custom call time would go wait get one never like
Topic 6
vegan got chip also tri taco burrito burger salsa mexican
Topic 7
best alway time love go recommend servic delici amaz friendli
Topic 8
coffe ice sandwich toast breakfast tea also cream egg sweet
Topic 9
realli beer park nice drink enjoy menu cocktail bar night
Topic 10
insid tabl distanc wear outsid mask seat dine covid staff
Topic 11
pork rib beef sandwich side meat bbq chees mac brisket

4.2. Main Analysis

4.2.1. Number of Reviews Mentioning COVID and COVID Cases

As noted in the methodology section, we placed each review into the highest-probability topic returned by the LDA model. This allows us to compute the number of daily mentions of each topic. In particular, given our interest in restaurant and customer responses to the COVID-19 pandemic, we calculate the daily number of reviews classified under topic 10 (the COVID topic). We then plot the smoothed number of daily mentions of the COVID topic (Topic 10) from our LDA analysis and compare this to the smoothed daily number of COVID cases in the United States from the CDC database (Centers for Disease Control and Prevention, 2021). In the plot below, we can see that COVID mentions do seem to mirror the COVID cases—with something of a lag. For example, 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.





4.2.2. COVID and Review Sentiment

Finally, in the main part of our analysis, we examine the link between COVID cases and the sentiments expressed in reviews. It is important to account for the hierarchical nature of our data—businesses within counties within states (Gelman & Hill, 2007). Therefore, we fit multilevel linear models predicting the sentiment of the reviews with varying intercepts (random effects) for businesses, counties, and states. Additionally, we included the lagged cumulative number of cases in a county as a predictor to examine the role played by local COVID in driving people's sentiments.

First, 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	

After establishing that restaurant customers are talking about COVID in their reviews and that COVID, at least using VADER sentiment polarity, appears to be driving at least some negative sentiment in reviews, we look to see how restaurant owners might reduce some of the negative reviews from COVID. We hypothesize that features of restaurants that are more COVID-safe might reduce the impact of COVID on restaurant ratings and the sentiments expressed in restaurant reviews. The Yelp businesses dataset contains information on whether the businesses have outdoor dining, delivery and takeout options, and a Drive-Thru. All of these features of businesses make it easier for people to avoid coming into contact with COVID. Again, given the hierarchical, or clustered, nature of our data, 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. This allows us to see both how restaurant features affect the sentiments expressed in the text of reviews and how restaurant features impact the effects of COVID cases. We expected that the interaction terms would be positive and significant, indicating that they reduce the negative impact of COVID cases on sentiment. Table 3 displays our results using VADER sentiment polarity scores, and Table 4 displays our results using TextBlob sentiment polarity scores.

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^{***}	-0.005^{***}	-0.002^{***}	-0.00004
	(0.001)	(0.001)	(0.001)	(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^{***}	0.629^{***}	0.632^{***}	0.514^{***}
	(0.012)	(0.019)	(0.015)	(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^{***}	0.001	0.001^{***}	0.002^{***}
	(0.0003)	(0.001)	(0.0004)	(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^{***}	0.248^{***}	0.257^{***}	0.206^{***}
	(0.006)	(0.009)	(0.007)	(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

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.

The coefficient estimates for the restaurant features are themselves interesting, as well, even aside from the interaction terms. Since our dataset of reviews is entirely from the COVID era, these estimates tell us the impact of these restaurant features on the sentiment of restaurant reviews during the pandemic. 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. On the other hand, outdoor dining's coefficients are positive and significant using both TextBlob and VADER sentiment polarity. Thus, it appears that having outdoor dining leads to more positive reviews.

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.

5. Conclusion

Throughout this paper, we have examined how the COVID-19 pandemic affected reviews of restaurants on Yelp. Our analysis was mainly directed at answering the question: how can restaurants best adapt to the COVID-19 pandemic, mitigating any increase in review negativity due to the spread of the coronavirus? We first showed that customers do appear to be talking about COVID in their reviews of restaurants, and the amount they talk about the pandemic appears to be related to the number of cases nationwide: an increase in cases nationally appears to be followed by an increase in reviews predominantly discussing the coronavirus. This analysis employed Latent Dirichlet Allocation to discern the most coherent topics in restaurant reviews, finding a clear COVID-related topic. Second, we utilized both VADER and TextBlob sentiment analyses to ascertain *how* people talked in their restaurant reviews during the pandemic–whether their reviews became more negative or positive. We found that as the local COVID caseload increased within counties, VADER sentiment polarity in restaurant reviews became more negative, while TextBlob sentiment polarity appeared to become more positive.

Finally, we showed that perhaps the best step that restaurants could take to prevent an increase in negativity in reviews was to provide an option for outdoor dining. Not only did an outdoor dining option have a direct association with the positivity of sentiment in reviews, but it also appeared to interact with the number of COVID cases (albeit insignificantly), potentially reducing the impact of a rise in COVID cases on the negativity of reviews. Other features of restaurants were somewhat mixed in their effects on reviews–though it seemed that both delivery and Drive-Thru options had clearly negative impacts on review sentiment. Businesses should carefully weigh their options in deciding whether to implement takeout or not. On one hand, takeout appears to be significantly negatively related to the sentiment expressed in a restaurant's reviews. On the other hand, it appears to have a positive interaction with COVID cases, reflecting a potential that takeout might mitigate some of the negative impacts in rising caseloads. Additionally, takeout might be one of the only alternatives for businesses choosing between shutdowns and takeout. In sum, our results suggest that restaurants confronting a rise in the number of cases due to the omicron variant should implement outdoor dining options to the extent possible and should carefully consider whether or not to allow for takeout.

While the COVID-19 pandemic posed new challenges for the restaurant industry, many restaurants adapted and were able to overcome some of the hardship. Our analysis tells us which steps taken by restaurants were most effective in reducing the negative effects of COVID on customer satisfaction–and importantly, which steps might have even been counterproductive. Thus, this paper provides a roadmap for what restaurants can do to adapt as we face a new variant.

References

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.