CPSC 599.27 - NLP Term Project Final Report

MATTHEW JARRAMS, University of Calgary, Canada, 30061191

In this Final report I will outline an overview of my work for last 3 months analyzing Yelp restaurant reviews in Edmonton. I will briefly introduce my topic and my methodology for handling my data and presents my results along with some discussion evaluating these results.

1 INTRODUCTION

Review websites like Yelp allow customers to rate and write about their experiences at businesses, providing valuable feedback for potential customers and business owners. However, most users only focus on quick overview numbers like average ratings, wasting potential data stored in customer descriptions. Motivated by this untapped data potential, my project used NLP methods and analysis techniques to extract information from review text and provide concise summaries to restaurant owners. I acquired a large dataset of restaurant reviews from Yelp[1] and using the methods in this article[2] extracted all restaurants in Edmonton. Leaving me with 68882 reviews from 2410 businesses to clean and use pre-processing methods to allow me to experiment with a variety of analysis strategies. This included topic modeling, sentiment analysis/classification, and visualization design. The goal was to present this information in a visually appealing manner to provide insights into restaurant owners' businesses and help them improve or maintain their current standards.

Previous work in this field has largely come from Yelp's data competitions, where participants submit projects that utilize the data and the best one is selected as the winner. These competitions have led to numerous analysis projects, including exploring sentiment and rating relationships [3], visualizing data with word clouds [4], and topic modelling [5]. However, the visualizations used in previous work tended to be fairly basic, and I plan to diversify my approach by creating more comprehensive dashboard-like visualizations for data results. Additionally, I observed a common trend of using different tokenization and text processing techniques, such as modifying stopword lists and exploring n-grams, to improve insights from reviews and enhance classifier and topic modeling results. I intend to further investigate these methods by incorporating different filtering strategies, such as restaurant categories and star ratings, and leveraging techniques discussed in our lectures this semester. For instance, I aim to compare classifier performance using various feature types and build upon existing work by incorporating insights from related articles.

2 MATERIALS & METHODS

2.1 Data & General Processing/Cleaning

In my data, I categorize reviews into two main classes: review ratings, which range from 1 to 5 stars representing the customer's experience (1: terrible, 2: poor, 3: OK, 4: good and 5: fantastic), and the main category of the restaurant (e.g., pizza restaurant, burger place, Chinese restaurant, etc.). Table 1 provides basic statistics about the star classes.

Key observation from this table are how the 4 & 5 star classes have almost 3 times more reviews than the 1 & 2 star classes. This was especially important during my classification where I experiment with different sample sizes from classes.

Table 1. Statistics for Star Classes

Star	Reviews	Unique Restaurants In Star Class	Length (words)	Blob Polarity	Vader Polarity
1	7586	1817	149	-0.06	-0.14
2	6652	1733	171	0.06	0.32
3	10420	2009	171	0.17	0.70
4	21377	2250	155	0.26	0.80
5	22846	2129	119	0.34	0.89

How I performed general cleaning and processing on my data was by:

- Extracting text from the reviews
- Removing numbers
- Converting all letters to lowercase
- Removing any special characters ('@', "/n", etc...)

These basic layers of cleaning ensured I had a clean set of tokenized words to use for my analysis.

2.2 Methods

From the various analysis methods I attempted in my notebooks I shortlisted a selection of results to include in this report. These include topic modelling, ScatterText visualization, classification/sentiment analysis and a chronological analysis of a restaurants average rating.

To categorize the reviews, I divided them into two groups: low-rated (1/2 stars) and high-rated (5 stars). I experimented with three topic modeling algorithms: LSI, LDA, and NMF. I found NMF the most successful and have included the LDA results in appendix B.1 for reference. From the topics in the results I observed that certain similar common words relating to customer dining experiences were dominating (e.g, table, service, place, etc...). In order to find more subtle insight behind the reviews I added these words to the removed stop words list and swithced to using NMF topic modelling. Using a similar technique to an article [5] that paired the topics found back to the restaurants I created a distribution of what topics reviews were roughly coming from for each restaurant.

For NMF the workflow consisted of:

- Create a TF-IDF representation of bigrams removing the same stop words as for the LDA model
- Normalize topic weights found to ensure all topic values for one review summed up to 1
- Pair topic weighted reviews back to the restaurants
- Calculate average values for each topic to show the distribution of topics among the reviews for an individual restaurant

One of my goals was to create a dashboard-style visualization for restaurant owners, which would provide a concise summary of reviews while also allowing access to the original raw data in one place.

To find this balance, I discovered a library called ScatterText from SpaCy[6]. Using a TF-IDF representation of uni-grams, bi-grams, and tri-grams with removed stopwords and common words as previously mentioned, I calculated scaled F-scores using positive and negative categories to generate scores for each bigram, representing their relative

frequency in both categories. These values were then plotted on a grid to create visualizations of the n-grams in relation to low and high-rated reviews.

For these results I took a particular category of restaurant (Pizza) and then created two groups low (1, 2 and 3 star) and high (4 and 5 star) rated. I show my results for these in the next section with some visualizations in Appendix B.2 as well as the HTML files submitted.

For classification, I utilized various approaches to compare the performance of different types of features. Additionally, I sought to extract meaningful insights from classification results that could be presented to restaurant owners in a simplified manner. Initially, I conducted a logistic regression on a collection of 1 and 2 star reviews (low category) and 5 star reviews (high category). To further investigate the important features influencing the classifier's predictions, I employed the SHAP (SHapley Additive exPlanations) library. I used a TF-IDF representation of the bi-grams with stopwords and common words removed as the input for the classifier.

Using a Random Forrest Classifier on two different class filters (1 vs. 5 and 2 vs. 4). I took 6000 samples from each rating class and then used 5 sets of of different features to compare.

- Manual text mined characteristics from the text
- TF-IDF representation
- Word2Vec model from scratch
- Pre-trianed Google News 300 Word2Vec model

For the manually selected features, I used the Vader library to find the number of positive/negative words and overall polarity scores and then extracted adverb and adjective counts, percentage of punctuation used and lastly the number of words relating to service, food and price. This included a selection of words commonly related to these three topics. For example, counting the number of times the word 'service' or 'cooked', 'cheap', etc...

The TF-IDF, word embeddings from scratch and the pre-trained Google News model were all run on bigrams with stopwords and other common words like 'service', 'table', etc... removed (see appendix A.1 for full list)

One of motivations for looking into classifying ratings by stars was due to some of the inaccuracies seen in just using plain sentiment scores from libraries like Blob and Vadar (Appendix B.3). Looking at table 1 above, the polarity does tend to follow an expected distribution (increasing towards the higher stars) however, there are quite a few inaccuracies. Which led me to investigate the individual reviews to look for clues as to why this could be happening. In appendix (A.2 & B.3) I have included some example reviews with their relative sentiment score and some investigation into sentiment scores. These example included reviews that write positively about service but negatively about food and some that use sarcasm. Both of these led to mislabelled sentiments.

Lastly, I recognized the importance of considering the date of reviews for restaurants. While analyzing topics and words related to different review categories is valuable, it may not capture recent changes or issues that the restaurant has addressed. Additionally, a restaurant with a high average rating might experience a sudden decline in reviews that

could be overlooked in visualizations like ScatterText or topic models.

For this analysis I used the following steps:

- Choose a restaurant
- Investigate reviews to find key words for this specific restaurant
- Using FreqDist to find the most common words to build manual topic groups (price, dishes, service, etc..)
- Sort reviews in chronological order
- Find cumulative averages (net rating scores, average rating and counts for reviews relating to these manually selected topics)
- Create visualizations, display specific reviews relating to dips or spikes with the key word topics attached

For more information regarding my methodlogy for calculations of scores please see Appendix B.4.

3 RESULTS

3.1 Topic Modelling

To see my analysis from LDA topic modelling please see Appendix B.1

Below are the results for the NMF topics, where the topics were easier to identify and I was able to show a plot of the different topics for the restaurant "LovePizza" where I give a breakdown of how the reviews for this restaurant are distributed (Figure 1). In appendix B.1 I give an example of two reviews with the relative weights to each topic to demonstrate how these distributions are found.

The main highlights from the NMF topics:

- Entire category just for Chinese food in the 1 star reviews
- Ability to extract topics of interest for business owners (Tables 2 and 3)
- Normalizing NMF topic weightings to present distributions of topics as percentages (figure 1)

From figure 1 below, the owner of LovePizza now knows around 70% of 5 star reviews are to do with first or multiple time visits and food and around 70% of 1 star reviews are to do with delivery and wait times.

Table 2. NMF: 1 Star Top Words

Topic	Top Bigram 1	Top Bigram 2	Top Bigram 3	Top Bigram 4	Top Bigram 5	Predicted Topic Labels
1	waited minutes	another minutes	took minutes	drive thru	minutes later	Wait Times / Fast Food
2	fried rice	spring rolls	deep fried	ginger beef	sweet sour	Chinese Cusine
3	first time	second time	time ordered	last time	twice first	Number of Visits
4	chicken staff	give instead	coming ill	staff coming	ill give	Hygiene

Table 3. NMF: 5 Star Top Words

Topic	Top Bigram 1	Top Bigram 2	Top Bigram 3	Top Bigram 4	Top Bigram 5	Predicted Topic Labels
1	first time	one best	one favourite	every time	spring rolls	Number of Visits/Food
2	friendly staff	super friendly	staff definitely	staff super	atmosphere friendly	Service / Atmosphere
3	love love	love absolutely	love best	awesome definitely	die buffet	Very Positive Experience
4	staff friendly,	friendly helpful	amazing staff	delicious staff	friendly knowledgeable	Staff / Food

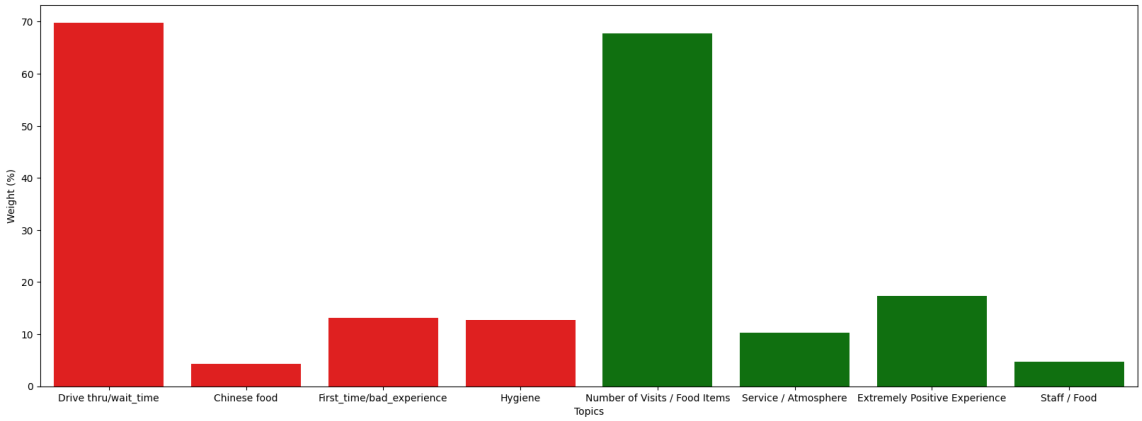

Fig. 1. LovePizza Review Distribution. Green indicates 5 star topics and red are the 1 star topics

3.2 ScatterText

For ScatterText, I provide guidance on interpreting the results and present alternative visualizations in Appendix B.2. However, to fully grasp the insights, it's best to explore the HTML files I've submitted.

From the bigrams I plotted on the grid (Figure 5 in appendix B.2 and PizzaScatterText.html) based on their frequency in each category using scaled F-scores. I extracted three key pieces of information:

- Bigrams in the bottom left ('tasty tomato', 'best donair' and 'right amount') all occur more frequently in the positive reviews.
- Bigrams in the top right like ('garlic toast', 'steak sandwich' and 'extra cheese') are more related to negative reviews.
- Bigrams in the top left occur frequently in both, these typically contained things like the names of restaurants ('boston pizza') and other words relating to pizza that could be negative or positive ('thin crust').

From my ScatterText analysis results I found:

- Bigrams offer better category separation compared to single words or trigrams
- Trigrams provide distinct category segregation, but lack specifics about individual dishes (See Appendix B.2)
- Bigrams strike a good balance between capturing customer experience and specific food topics, making it possible for owners to click on dish names and view reviews related to specific food types.

.

3.3 Classification

For my classification results I have included a table below with all of the results from the random forest classifier with the various different features used.
For Notes about data and models

- Data had 6000 samples from each star class (12000 total reviews)
- model parameters (n_estimators = 1000, min _samples_split = 2, min_samples_leaf = 1)
- Data Split: 75/25 (with stratification)

Table 4. Random Forrest Models (1 vs. 5 star classifications)

Features Type and Class	Precision	Recall	F-1 Score	Overall Model Accuracy
Text Minded Features (1 star)	0.92	0.91	0.92	0.92
Text Minded Features (5 Star)	0.92	0.93	0.92	0.92
TF-IDF (1 star)	0.82	0.69	0.75	0.77
TF-IDF (5 star)	0.73	0.85	0.79	0.77
Word2Vec - Scratch (1 star)	0.72	0.79	0.79	0.74
Word2Vec - Scratch (5 star)	0.76	0.69	0.73	0.74
Word2Vec - Google-News-300 (1 Star)	0.92	0.93	0.93	0.93
Word2Vec - Google-News-300 (5 Star)	0.93	0.92	0.93	0.93

Table 5. Random Forrest Models (2 vs. 4 star classifications)

Features Type and Class	Precision	Recall	F-1 Score	Overall Model Accuracy
Text Minded Features (2 star)	0.78	0.77	0.77	0.78
Text Minded Features (4 Star)	0.79	0.80	0.79	0.78
TF-IDF (2 star)	0.70	0.74	0.72	0.71
TF-IDF (4 star)	0.72	0.68	0.70	0.71
Word2Vec - Scratch (2 star)	0.64	0.66	0.65	0.64
Word2Vec - Scratch (4 star)	0.65	0.63	0.64	0.64
Word2Vec - Google-News-300 (2 Star)	0.82	0.82	0.82	0.82
Word2Vec - Google-News-300 (4 Star)	0.82	0.82	0.82	0.82

The key takeaways from these classifier results (Tables 4 and 5):

- Relative decrease in accuracy in each category for classifying 2 vs. 4 star reviews compared to the 1 vs. 5 star. (See Appendix B.3 for side investigation)
- High success of manually extracting features from text (positive words, punctuation, number of key words in certain topics, etc...) compared to simple word vectors and embeddings (Word2Vec scratch)
- How accurate the pre-trained Word2Vec model performed with minimal fine-tuning compared to the other text representation features.

While the academic approach above may not be suitable for a business dashboard, I used the SHAP library to investigate feature importance in logistic regression. The table below shows the results of running the model on low-rated (1/2 stars) and high-rated (5 stars), for my features I used a TF-IDF vectorizer of bigrams.

Table 6. Logistic Regression (low (1/2 star) vs. high (5 star) accuracy report)

Class	Precision	Recall	F-1 Score	Overall Accuracy
Low rated	0.95	0.33	0.49	0.74
High rated	0.70	0.99	0.82	0.74

Since I did not do any sampling I had twice as many high rated reviews than low rated reviews leading to the model bias towards the high rated reviews. Despite this bias highlighting the importance of sampling, the model still provides interesting insight into what bigrams are more related with positive and negative reviews.

From the visualizations below (Figure 2) the key results are:

- Good customer service and having friendly staff was linked with positive reviews
- Reviews not containing "reasonably priced" were more likely positive
- When gluten free was in the review it had more of a positive impact than something like "deep fired" which was more negative

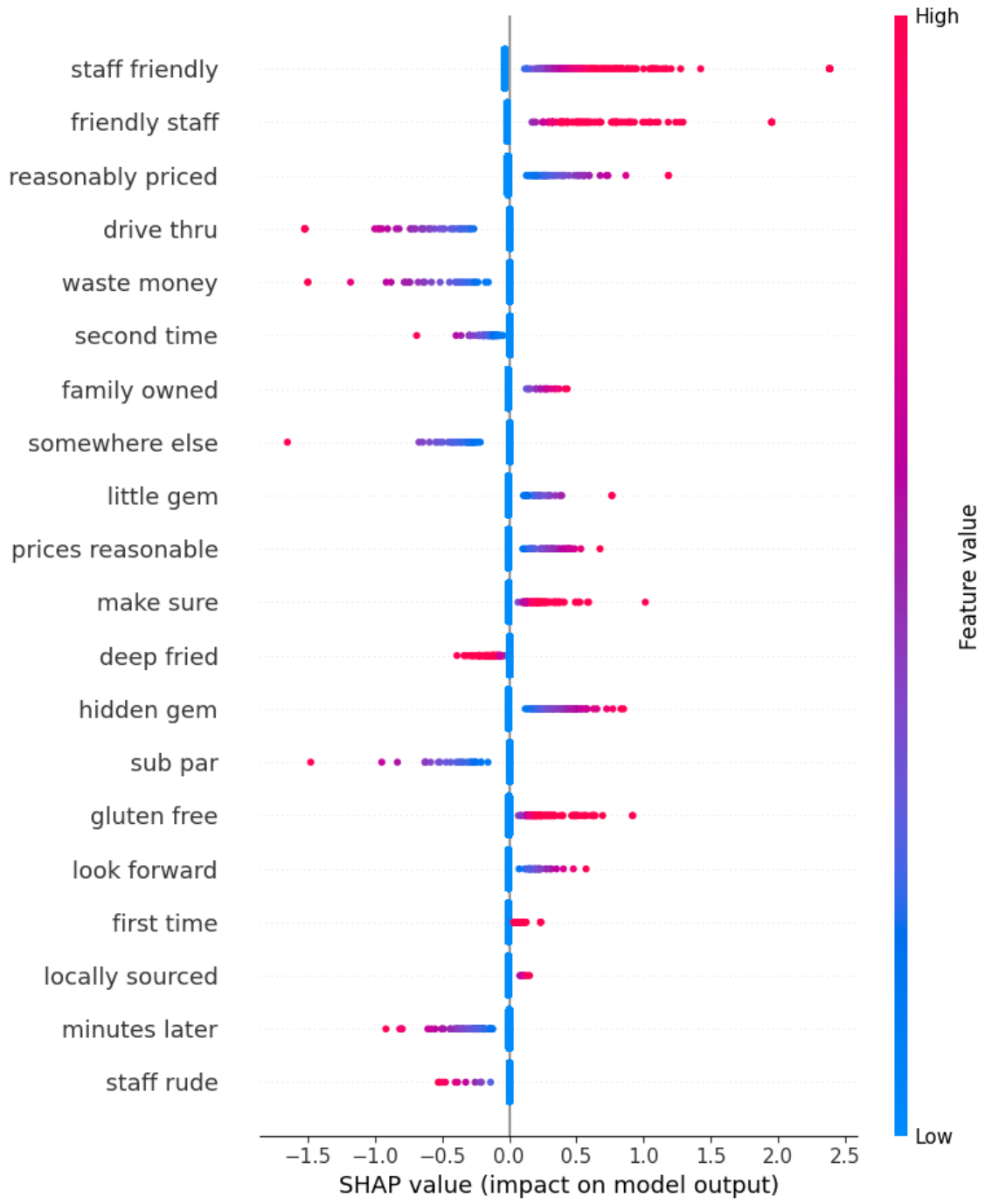

Fig. 2. SHAP Plot of most important TF-IDF bigrams for classifying 1/2 star vs. 5 star reviews

3.4 Time Analysis

Lastly, my goal was to provide a restaurant with not only insights into what there reviews were about but to also give a time frame. I choose "Chianti Cafe & Restaurant - Old Strathcona" (121 reviews with an avg. rating of 3.5) to focus on. The chosen topics to focus on were food (popular Italian dishes like pasta), service and price (words like Monday and Tuesday included as these are nights with special deals). See appendix B.4 for a full list of key words chosen.
Below are some resulting visualizations.

Table 7. Chianti Topic Distribution

Topic	Net Rating Score	Positive (%)	Negative (%)
Dishes	12.40	66	34
Service	5.79	63	37
Price	4.13	56	44
Other	18.18	63	37

Table 8. Reviews from Chianti's biggest dip in average rating

Date	Stars	Topic
2010-10-06	4	Dishes
2011-01-19	2	Service
2011-01-29	2	Other

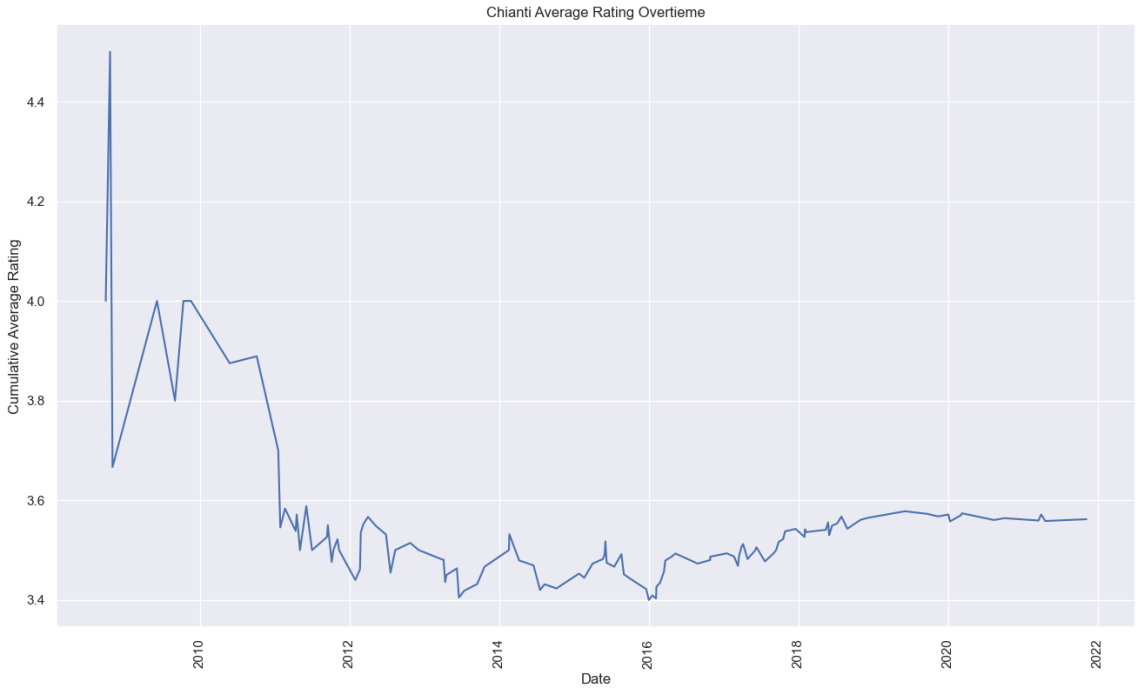

Fig. 3. Chianti's average rating overtime

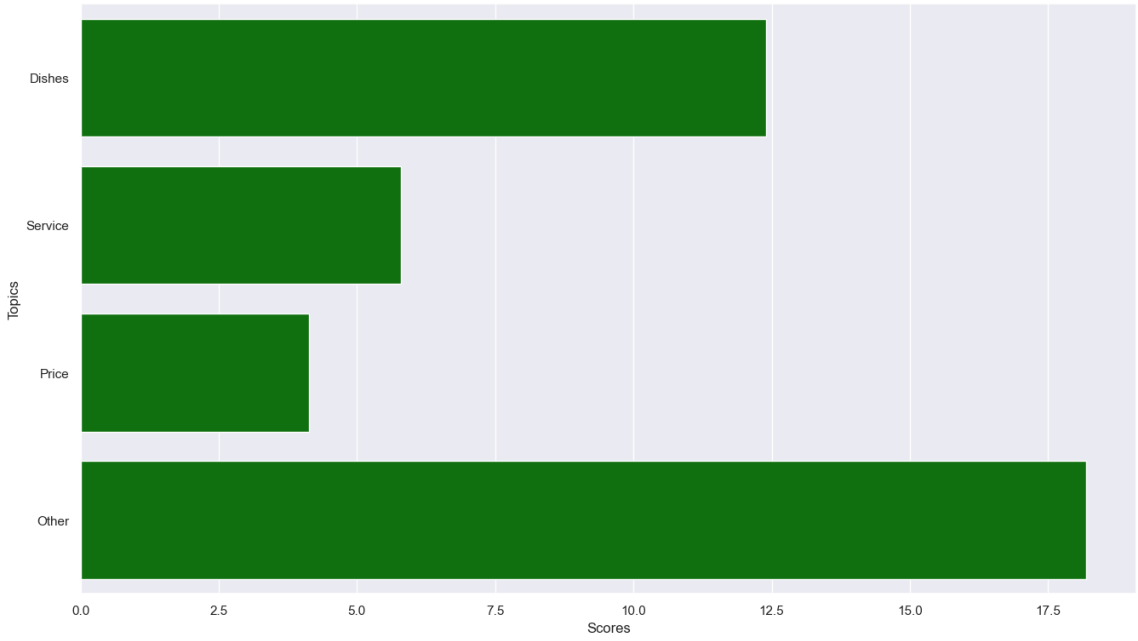

Fig. 4. Chianti specific topic's impact on score. Calculated by finding the difference between that topic's positive & negative reviews divided by the sum of all reviews. Result is the net impact that topic has had on the average rating expressed as a %.

Highlights from the above graphs and tables:

Line Plot (Fig 3 / Table 7)

- Biggest dip in Chianti's average rating (first 5 reviews omitted since any rating will cause large fluctuations with small sample size) came around the end of 2010 and beginning of 2011
- During there largest dip they had 4 star dish review and then two 2 star ratings one about service and the other had no topic
- In recent years (2020-2022) their average rating has stayed fairly constant, causing no need for drastic changes.

Topic Impacts (Fig 4 / Table 8)

- In this case Chianti's chosen topics all have positive impacts
- Reviews relating to the dishes they serve are the most positive
- Higher number of negative reviews related to reviews mentioning price
- High number in the "other" category imply the chosen key words will need more refining to better capture the topics of all reviews

4 DISCUSSION AND CONCLUSION

To evaluate the success of these results in how well this analysis could be utilized by business owners I will provide interpretation to some of the results to suggest some possible outcomes and use cases.

Topic modelling revealed that NMF results provided a better overview of review topics compared to LDA topics. Normalizing the values using NMF allowed for more readable percentage values, helping business owners identify areas for improvement. By combining topic distributions with ScatterText plots, a dashboard could be created for restaurant owners to analyze reviews with specific keywords of interest. However, both topic modelling and ScatterText have limitations, such as my choice to remove common words. This may have impacted the relative context around these removed words. Further pre-processing could improve topic labelling and ngram representations in ScatterText plots in the future.
ScatterText's strength lies in categorizing words or bigrams into positive or negative frequency levels and providing an interactive and accessible way to access raw data. Using ScatterText for a Pizza restaurant, I identified that Italian Sausages and Butter Chicken (as pizza toppings) were associated with positive reviews, while "extra cheese" and "garlic toast" were more frequent in negative reviews. Instead of manually reviewing all their feedback, restaurant owners can request a ScatterText plot to analyze reviews more efficiently. This tool can be invaluable for restaurant owners seeking to delve deeper into individual review content in a streamlined manner.

My classifier results provided an academic investigation, showcasing topics discussed in class. Although some byproducts of the classifications provided some interesting indications to what factors drive positive and negative reviews. One of the most intriguing results was the high accuracy of the Google News 300 Word2Vec model with minimal fine-tuning. It outperformed other methods in both 1 vs. 5 and 2 vs. 4 star classifications, including the TF-IDF representations and the Word2Vec model trained from scratch. The pre-trained model's word embeddings captured the contextual usage of words across different reviews, making it more powerful than TF-IDF which focuses on individual word frequency. The larger amount of data and training time of the pre-trained model likely contributed to its overall better performance.

It was also noteworthy that the accuracy decreased in all classifiers when comparing 2 and 4 stars to 1 and 5 stars. This is expected as 2 and 4 star reviews are closer in rating and are less likely to have extreme emotions compared to the polar opposite 1 and 5 star reviews. In appendix B.4 I have included some supporting tables that show the differences between 1/5 and 2/4 reviews.

My classifier results were fairly similar to this analysis [4], in that we were both able to find topics to fit restaurants. However, in my analysis I used bigrams as opposed to single words and the notebooks results are stronger than mine were for single words. This could be due to leaving certain words in that I removed or more detailed cleaning and processing. Taking all classification results together the keys for business owners are:

- Taking note of bigrams in the SHAP plot to that show common words that relate to more positive reviews and negative reviews
- Looking at what absences of certain bigrams have an effect.
  - "reasonable priced" has a positive impact when not present indicating a possible pattern that the price could be a saving grace of the restaurant for the customer. Whilst the food and service are negative overall the restaurant is not that expensive.
- Offering accommodating food options like "gluten free" and less "deep fried" food could increase ratings as "gluten free" is more related to positive ratings whilst "deep fried" tended to be more negative.

To enhance the analysis, further fine-tuning of models and vector/embedding representations would be beneficial to assess their accuracy. In comparison to the analysis [4], my results using TF-IDF were not as accurate, particularly in

the 1 vs. 5 star reviews. However, this discrepancy could be attributed to the use of a different classifier. Had I employed a MultinomialNB algorithm, the results might have improved.

Finally, while examining a restaurant's average rating over time, I found potential improvements to the method. Using personalized keywords related to the business yielded good results, although there was a high percentage of reviews categorized as 'other'. Despite this limitation, the advantage is that business owners can still see which review topics (such as food, service, or price) are impacting their ratings more heavily, making decision-making easier. This is particularly helpful as sifting through all reviews on Yelp's website can be challenging. Combining this information with identified review topics could create a dashboard for restaurant owners to quickly identify recent trends in review ratings and associated topics.

The list below summarizes the limitations and improvements that could be made to all results.

- Greater investigations into extra stop word removal
- Different filtering queries (which categories to drop or to focus on)
- Creating equal distributions of categories and star classes
- Running results on different ngrams (trigrams)
- Investigating specific word detection like SkipTheDishes and UberEats could have been beneficial for extracting reviews specific to these categories
- Analyzing time in groups of 10 to avoid early reviews dominating fluctuations in average rating

Despite these limitations, the results presented were successful in the following areas:

- Possible dashboard content based on NLP statistics as opposed to popular summary numbers seen on Yelp and other typical dashboards.
- Comparing the performance of a variety of different feature creation strategies based on the review text for classification in order to predict star rating
- Providing indications on which words drive positive and negative reviews

Lastly, all results to some degree contribute to summarizing the lengthy text reviews provided to business and provide a much more appealing resource than reading them one by one.

Looking to future work, the following are some possible directions,

- Test models on data from one of the other 10 metropolitan areas in the dataset to see if my models were more bias to Edmonton restaurants or not
- Used my Word2Vec model to create t-SNE plots (plot reviews into space and look for clusters)
- Analyze customer impact (peer feedback from presentation)
  - Using the customer JSON file, future analysis could be done in finding frequent Yelp users and applying increased weightings to reviews they make.

These ideas would be an interesting direction to take this project and provide good places to start for building on my results.

5 REFERENCES

(1) Yelp. 2023. Yelp open dataset. (2023). Retrieved January 28, 2023 from https://www.yelp.com/dataset

(2) George Hou. 2020. Converting yelp dataset to CSV using pandas. (February 2020). Retrieved January 26, 2023 from https://towardsdatascience.com/converting-yelp-dataset-to-csv-using-pandas-2a4c8f03bd88

(3) Louise Morin. 2022. Using NLP to extract insights from your customers' reviews. (May 2022). Retrieved January 26, 2023 from https://www.artefact.com/blog/using-nlp-to-extract-quick-and-valuable-insights-from-your-customers-reviews/

(4) Zhenyufan. 2019. NLP for Yelp Reviews. (February 2019). Retrieved January 26, 2023 from https://www.kaggle.com/code/zhenyufan/nlp-for-yelp-reviews

(5) Ankur Vishwakarma. 2018. NLP analysis of yelp restaurant reviews. (March 2018). Retrieved January 25, 2023 from https://medium.com/@Vishwacorp/nlp-analysis-of-yelp-restaurant-reviews-30b3d0e424a6

(6) Jason Kessler. SCATTERTEXT · spacy universe. Retrieved April 16, 2023 from https://spacy.io/universe/project/scattertext

(7) JasonKessler. 2017. Beautiful visualizations of how language differs among document types. (July 2017). Retrieved April 16, 2023 from https://github.com/JasonKessler/scattertext

(8) George Hou. 2020. Yelp_dataset/yelp_rv_nlp_scattertext.ipynb at master · Gyhou/yelp_dataset. (February 2020). Retrieved April 16, 2023 from https://github.com/gyhou/yelp_dataset/blob/master/notebooks/yelp_RV_nlp_scattertext.ipynb

(9) Jason S. Kessler. 2018. Natural Language Visualization With Scattertext. (April 2018). Retrieved April 16, 2023 from https://nbviewer.org/github/JasonKessler/GlobalAI2018/blob/master/notebook/Scaled-F-Score-Explanation.ipynb

(10) Scott Lundberg. 2017. Sentiment analysis with logistic regression. (September 2017). Retrieved April 16, 2023 from https://slundberg.github.io/shap/notebooks/linear_explainer/Sentiment%20Analysis%20with%20Logistic%20Regression.html

(11) Zhi Li. 2019. A beginner's guide to word embedding with Gensim word2vec model. (June 2019). Retrieved April 16, 2023 from https://towardsdatascience.com/a-beginners-guide-to-word-embedding-with-gensim-word2vec-model-5970fa56cc92#:~:text=We%20can%20train%20the%20genism,and%20the%20default%20is%20100

A DATA INFORMATION

A.1 Extra Information about Methods

Extra words removed for certain processing methods:
"go", "recommend", "like", "back", "good", "get", "even", "food", "place", "us", "order", "would", "restaurant", "meal", "got", "came", "could", "went', "said", "table", "asked", 'really', "never", "worst", "ever", "food', 'service', "place, "us", "order", "would", "restaurant", "meal", "got", "came", "could", "went", "said", "table", "asked", "really", "never", "worst", "ever", "highly" , "delicious", "definitely", "better", "one", "always", "well", "best", "edmonton", "perfect", "perfectly", "great", "awesome", "nice", "bad", "top", "every", "favorite", "favourite", "super", "nothing", "also", "try", "amazing", "love", "keep", "coming","terrible", "disappointed", "everything", "excellent", "fantastic"

A.2 Inconsistent Polarity Scores With Review Rating

Mislabelled 1 star as 5 review:

- **Review:** "Just waited 25 min in drive through with one vehicle in front of me worst service I have ever received at McDonald's and there have been some bad ones"
- **Rating:** 5
- **Blob Polarity:** -0.85
- **Vader Polarity:** -0.8225

Negative words used in positive way

- **Review:** "Never had a bad experience. Avocado shake, viet coffee, pho, vermicelli are my favs."
- **Rating:** 5
- **Blob Polarity:** -0.70
- **Vader Polarity:** 0.28

Negative words used in positive way and not a review of the restaurant

- **Review:** "Can someone please confirm if this restaurant has permanently closed it's doors? This is terrible news!"
- **Rating:** 5
- **Blob Polarity:** -0.55
- **Vader Polarity:** 0.20

Mixed experiences with restaurant aspects

- **Review:** "Customer was talking about me in another language. Although the food was good I do not recommend them for their insufficient customer service."
- **Rating:** 1
- **Blob Polarity:** 0.7
- **Vader Polarity:** 0.20

B SUPPORTING VISUALIZATIONS

B.1 Topic Modelling

LDA

Key observations from the LDA topic modelling in tables 9 and 10:

- Bigram 'first time' and references to visit frequency were mentioned the most in topics
- 5-star topics were harder to label with a general summary than 1-star topics.
- Both sets of topics were not as helpful as NMF

in tables 2 and 3." For more visualizations of LDA topics, refer to appendix B.1 or my Python notebooks.

Table 9. LDA: 1 Star Top Words

Topic	Top Bigram 1	Top Bigram 2	Top Bigram 3	Top Bigram 4	Top Bigram 5	Predicted Topic Labels
1	one star	skip dishes	waited minutes	every time	deep fried	Delivery / Fast Food
2	first time	last night	waste money	stay away	save money	Bad Experience / Money
3	first time	minutes later	long time	another minutes	last time	Waiting Times

Table 10. LDA: 5 Star Top Words

Topic	Top Bigram 1	Top Bigram 2	Top Bigram 3	Top Bigram 4	Top Bigram 5	Predicted Topic Labels
1	every time	one best	first time	bubble tea	fish chips	Food / Consistently Good
2	cactus club	first time	pulled pork	friendly staff	little bit	Cactus Club / Staff
3	gluten free	make sure	ice cream	deep fried	super friendly	Food Options / Types of Food
4	love love	always friendly	first time	sherwood park	staff always	Sherwoord Park Resturants
5	spring rolls	butter chicken	first time	one best	green onion	Asian Cusine

Example Reviews With Topic Distribution (NMF)

- **1 Star Review**: "Ordering online looked like it would be easy enough. Ordered 2 pizzas yesterday and online said 60 minutes for delivery. 60 minutes passes and I get a message that the driver will be there in 2 minutes. I go down to the front door to greet the driver and NO driver. I called LovePizza and they tell me that the driver was at my place for 15 minutes. I called BS on that statement. They then had the audacity to tell me the driver will return in 25 minutes. They also claimed they couldn't call me on the cell number I was calling them on. They had my email and cell phone number as confirmed on the phone call. Place is run by lying morons. I won't bother with them again with the amount of better run establishments in Edmonton."
- **Topic 0:** drive thru/wait time = 0.9968
- **Topic 1:** Chinese food = 0.0000
- **Topic 2:** first time/bad experience = 0.0000
- **Topic 3:** hygiene = 0.0032
- **5 Star Review**: "We've already been her twice this month. It's pretty quick, reasonably priced, the staff are friendly, and the most importantly the food is fresh and delish. We tried one of their signature pizzas the first time and they came together quite nicely. I will advise that the Peaches Cream is sweet enough to be a desert, but it is so good. The second time around we decided to design our own and there are so may great topping options to please any taste buds. Their (conveyor belt) oven does a perfect job at cooking the pizza to a nice crisp. The portions are large enough that half a pizza might be enough for one sitting or to share. We will definitely be back again to try another creation and maybe a dessert pizza while were at it."
- **Topic 0:**Number of Visits / Food = 0.2608
- **Topic 1:** Service / Atmosphere = 0.0000
- **Topic 2:** Topic 2: Very Positive Time = 0.0000
- **Topic 3:** Staff / Food = 0.7392

PyVisuals for LDA

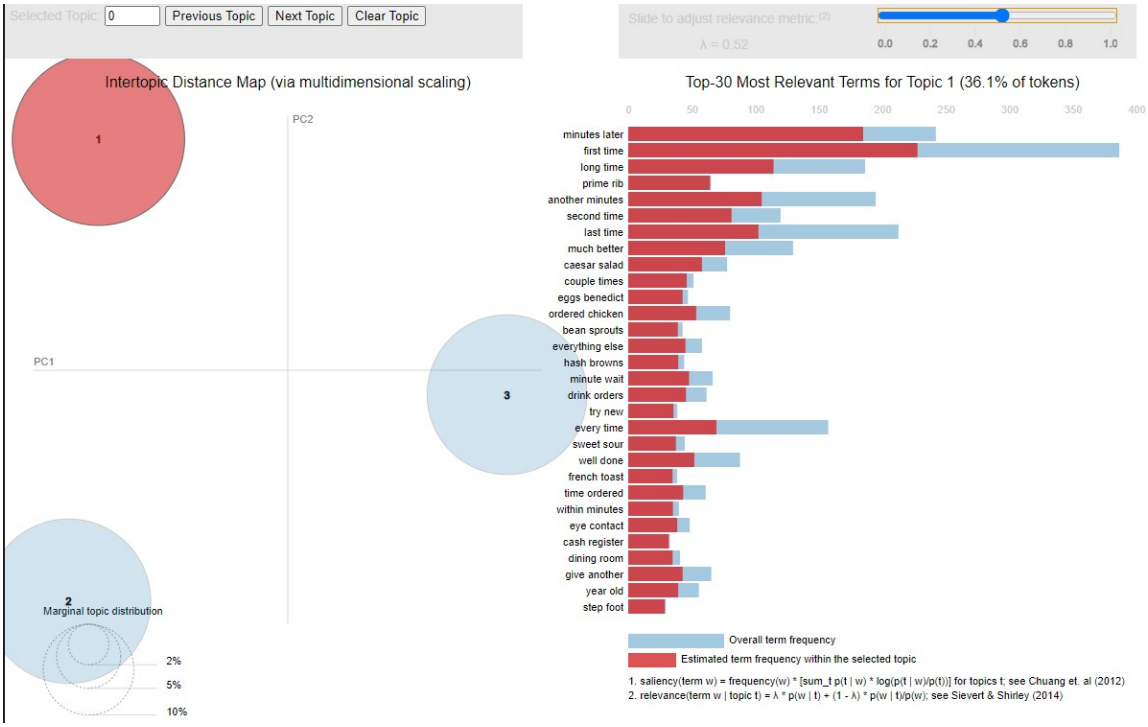

Fig. 5. 1 Star LDA Visual. 3 Topics all spaced out evenly representing Food, wait times and bad experiences

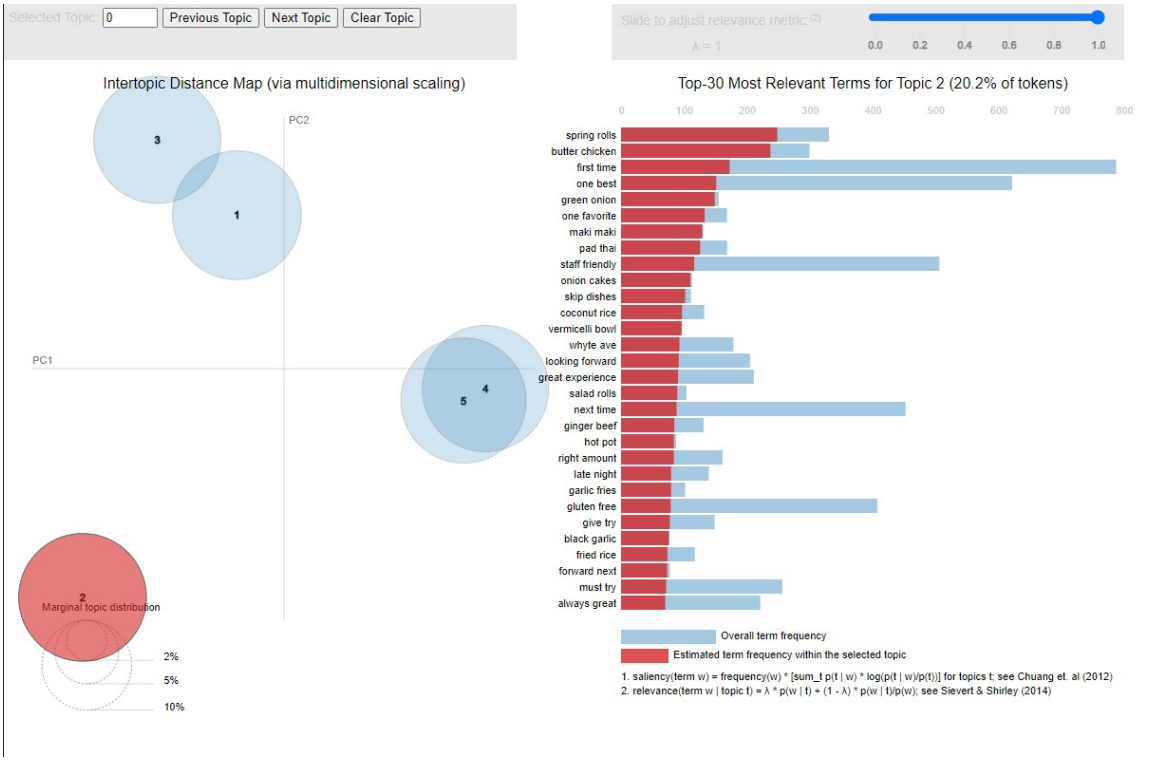

Fig. 6. 5 Star LDA Visual. 5 Topics with less separation than 1 star topic. Topics represent specific locations, great experiences and consistency

B.2 ScatterText

On the next page there are supporting images for ScatterText, showing some of the features and specifically figure 7 which is relevant to the results in 3.2.

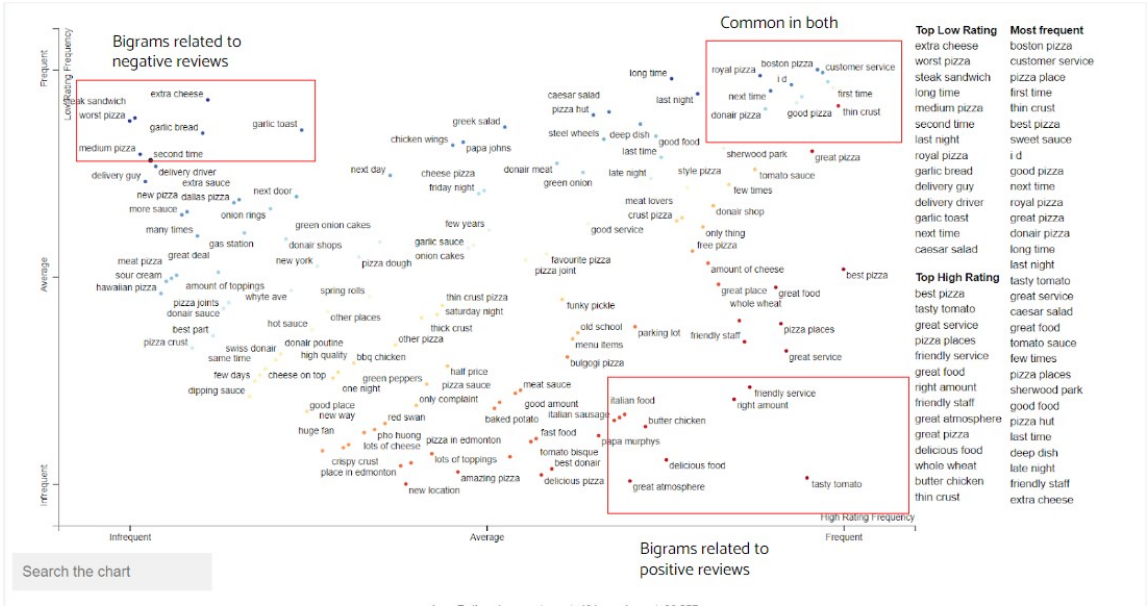

Fig. 7. ScatterText Dashboard of Pizza Resturant Bigrams.

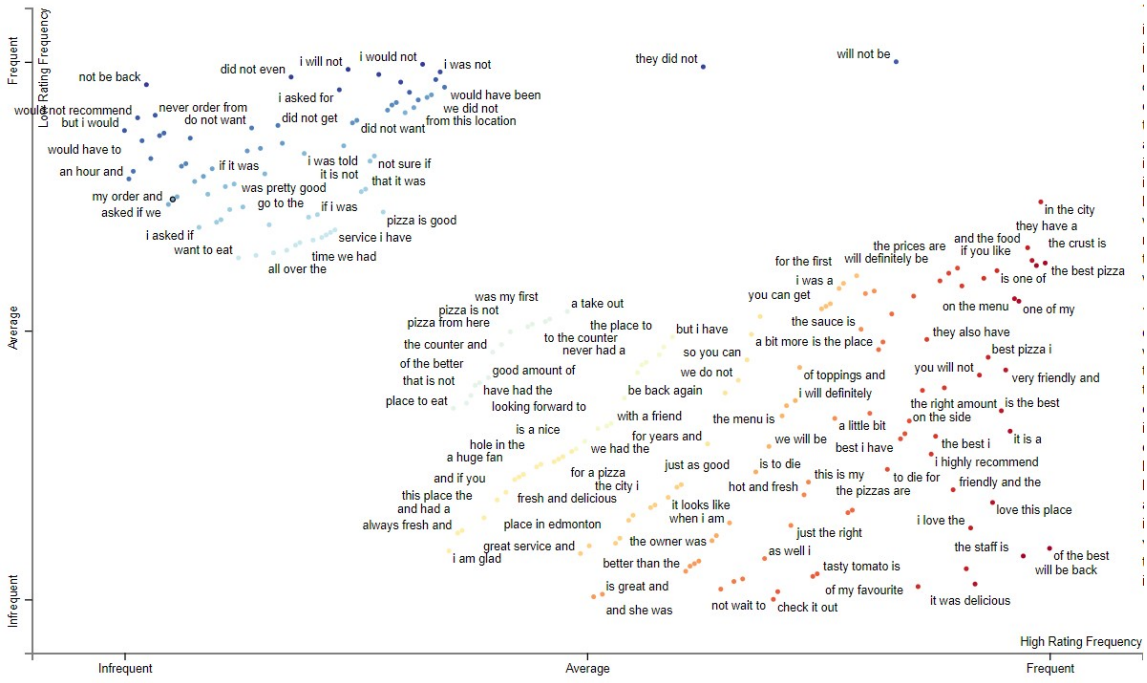

Fig. 8. ScatterText for pizza restaurants (Trigrams)

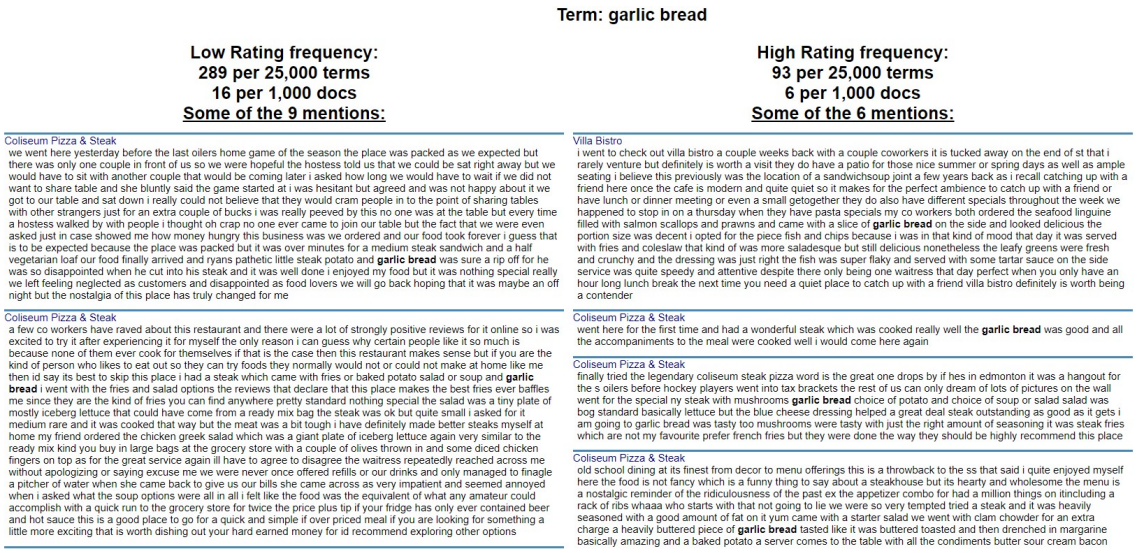

Fig. 9. ScatterText view for individual word selected. In this image "garlic bread" from Fig 7 was clicked.

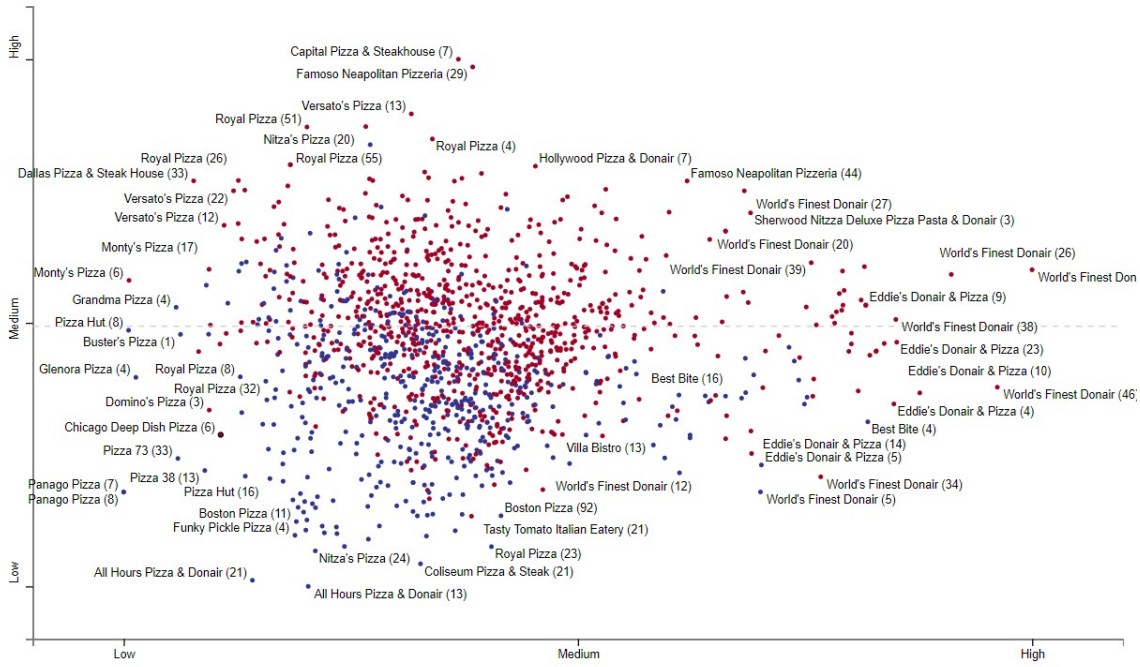

Fig. 10. Using ScatterText to place reviews into space. Experimented with maping reviews into space with a TF-IDF vectorizer. Results were not concrete enough to include in report.

B.3 Classification

Section includes information the SHAP library provides on individual data points (with an example). Along with a side investigation into how well Vader and Blob classify ratings using their sentiment scores. This provides the pretext for why I looked into classifying star rating and wanted to compare different strategies.

Individual Reviews

- **1 Star review:** "The tuna sashimi they gave me in my take out order was inedible. Such a waste of money. Well fool me once. I will not be back! You can't serve rotten fish to people! Additionally the server that took my order was way too sick to be at work. Really unappetizing and not appropriate when working in a restaurant. "

Fig. 11. Bigrams impacting this 1 star review (SHAP visulization)

Explanation of Blob and Vader Inaccuracy

The below tables highlight the inaccuracies and differences in using sentiment scores to classify ratings.

I used two calculations to fill the table below. The first was finding how many 1 and 2 star reviews had sentiments above 0 and how many 4 and 5 star reviews had sentiments below 0. Dividing these values by the total number of reviews in that star class gave me the percentage of the time the sentiment agreed with the rating (negative sentiment: 1/2 star review or positive sentiment: 4/5 star review).

To calculate the accuracy (sentiment and star agree) I assigned sentiment ranges to each rating.

- 5 star: sentiment score > 0.5
- 4 star: sentiment score < 0.5 but > 0
- 2 star: sentiment score < 0 but > -0.5
- 1 star: sentiment score < -0.5

I then divide these values by their relative class size to find the percentage of the time the sentiment is accurate in classifying the review.

Table 11. Vader and Blob performance classifying reviews using ranges.

Star	Sentiment & Star Disagree (Blob)	Sentiment & Star Disagree (Vader)	Sentiment & Star Agree (Blob)	Sentiment & Star Agree (Vader)
1	41%	5%	3%	82%
2	70%	67%	29%	13%
4	2%	0%	91%	4%
5	1%	1%	15%	96%

Key takeaways from this table

- Vader very strong at classifying 1 and 5 star reviews
- Blob very weak in classifying 1,2 and 5 star reviews
- Both Vader and Blob have high numbers of 2 star reviews that have positive sentiment scores

What these results indicate is that Vader typically has higher and lower scores which explain the higher accuracy in answering 1 and 5 star reviews in the correct range. Blob seems to assign values in and around 0-0.5 which is why it is stronger in classifying 4 star reviews. Overall these results highlighted the need to investigate my own classification methods instead of relying on the Vader and Blob libraries.

Comparison of 1/5 star and 2/4 star reviews

Some context behind the differences between the 2/4 star features vs. the 1/5 star features is shown in the below table. This supports the idea that 1 and 5 star reviews tend to have higher levels of polarity in both positive and negative directions than the more moderate 2 and 4 star ratings.

Table 12. Feature comparison between 2/4 star ratings and the 1/5 star ratings

Data Class	Positive Word Count	Positive Word Count	Polarity Variance
1/5 Star	23882	6440	0.52
2/4 Star	25286	4550	0.32

From the table the main differences to not are the fewer number of negative words in the 2/4 star and the lower variance in polarity score. Despite, having more positive words I believe this was the reason for the lower accuracy between the 1/5 star text mined feature classifier and the 2/4 star. The classifier likely had an easier job distinguishing 1 and 5 star reviews due to the higher variance in the polarity scores along with the higher number of negative words.

B.4 Time Analysis

Chianti Key Words Selected for Chosen Topics

Dishes: "pasta", "sauce", "italian", "dishes", "pastas","salad", "delicious", "fresh", "cream", "chicken", "dish", "spicy", "cooked", "tasty", "dessert", "frenzy"

Service: "waitress", "minutes", "service", "staff", "freindly", "served"

Price: "price", "cheap", "expensive", "portion", "special", "half", "reasonably", "size", "deal", "specials", "tuesday", "mondays", "monday", "tuesdays"

Explanations on Calculations

- **Topic Labelling:** Count the number of key word occurrences in a review and assign the topic with the most occurrences.
  - If no occurrences or less than 2 in each category label the review in the other category.
- **Cumulative Average:** Sum all review ratings to that date and divide by the count of all reviews to that data
  - This method of calculation meant reviews made earlier in a restaurants time will have more waiting in fluctuations in the average rating.
- **Topic Net Rating Score:** Sum all positive and negative reviews that are assigned that particular topic. Then divide the difference between these two sums by the total number of reviews that are labelled with this topic to get a percentage value.
  - If the value is negative, this topic has a negative impact on the restaurants average rating
  - If the value is positive, the topic has a positive impact on the restaurants average rating