Leveraging Online Reviews for Evaluating Business Performance

Anshul Kumar, Jaanvi Malik, Isidro Quevedo, Ananth Narne

May 2, 2025

Abstract

This project presents a comprehensive framework for analyzing and predicting business performance through customer reviews and combines rule-based methods, large language models (LLMs), supervised and unsupervised learning, and a recommendation system. We began with a comparative analysis of rule-based sentiment analysis tools to classify review sentiment, which included VADER and TextBlob, and found that they were inconsistent when tested against a gold standard set of 30 manually annotated reviews. We then moved to a transformer-based model called bart-large-mnli and used its zero-shot sentiment classification ability, which resulted in performance improvement, but it failed to label "Neutral" reviews appropriately. To improve reliability, we then used the ChatGPT API, particularly GPT 4o, to label the reviews, which proved to be more consistent when tested on the gold standard.

We calculated year-over-year deltas (2018-2019 and 2019-2020) for sentiment score and other metrics of business engagement, and used them to train the XGBoost classifier that predicts whether business engagement in 2021 increased, decreased, or remained stable. The model achieved an accuracy of 74% and macro F1-score of 0.58 on the test set, with strong precision on stable businesses and moderate performance on minority classes.

In parallel, a K-Means clustering model grouped businesses based on sentiment, average rating, and review volume, showing distinct segments of participation. A 3-D visualization and spatial mapping across California further illustrated regional patterns. Finally, we built a location-aware recommendation system using KNN and cosine similarity, combining TF-IDF review vectors, metadata, and sentiment. The system identifies similar businesses within a specified radius and provides actionable improvement based on POS-tagged key differences. Our results indicate that the techniques we have used and applied throughout the project show promising results in evaluating business performance over a wide array of parameters, including customer sentiment, which can give businesses significant leverage in understanding customer experience and making data-driven strategic decisions for sustained engagement and growth.

1 Introduction and Background

1.1 Introduction

As the Internet and technology continue to become part of the daily lives of people, online reviews have been more commonly used and increasing rapidly, becoming a source of information rich in customer concerns, sentiments, and opinions (Bi et al., 2019). Online reviews

are not only insightful, but have also proven to be a publicly available, large, easy-to-collect, and low-cost resource compared to other data collection methods. Online reviews have been successfully used for multiple kinds of decision analysis, such as customer satisfaction modeling, customer preference analysis, etc. (Bi et al., 2019). Previous studies have explored sentiment analysis and text mining techniques to extract product attributes and consumer sentiment. In addition to geographical proximity and similarity, these methods have been used to identify competitors within the same category (Gao et al., 2018). Analyzing online reviews can provide businesses with valuable insights into customer expectations, market trends, and competitive positioning. However, given that online reviews are unstructured, it is usually difficult to use a model that extracts structured insights from them. To address this challenge, studies have explored text mining and sentiment analysis to convert unstructured data into meaningful patterns (Guerreiro and Rita, 2020) We aim to examine how online reviews can be used as a tool to analyze business performance, providing insights into consumer sentiment, market trends, and competitive dynamics.

1.2 Background

The origin of reviews and leaving customer feedback first started back in 1999 (Shahare, 2022), making it a fairly new concept. Prior to the concept of online reviews, information about a product or service was spread by word of mouth and by credible, authoritative figures around the world (Sprague, 2025). For example, if there was a new restaurant that opened up or a new product that had been released, people would only be able to know about that if a close friend or family member of theirs had also gone there or if they knew someone who purchased that product. This concept changed after the emergence of the Internet. The big players in the field, commonly referred to as the "Big Five": Yelp, Amazon, Google, Facebook, and TripAdvisor, have all developed significant clout and influenced the evolution of online reviews" (Shahare, 2022). These companies were the first to notice the vast potential that reviews could serve and rushed to spread these to the mass markets.

The first online review was actually done in order to save paper and remove the tedious process of collecting written feedback from customers (Jones, 2018). This had initially started as a way to prevent a complicated process and had become an instant success due to the reliability and ease of interpreting online reviews. This was the era of the Internet and the use of technology was spreading fast all around the world. Businesses had started to roll out the concept of online reviews through the Internet, which was new and fascinating at the time for the general public. Fast forward 16 years, online reviews are now a major part of everyone's lives and many cannot order a product or visit a business without looking at reviews first. In fact, "72% of people say online reviews help them establish trust in a business" (Nowak, 2022). Online reviews are not just helpful for the customers; rather, the businesses themselves benefit greatly as well and the relationship between the two is considered to be mutualistic.

Even though there were not many studies done on online reviews because of their recent introduction to the market, many businesses had jumped into the world of online reviews and utilized what they had to offer in order to potentially boost their performance and increase their customer base. There were many positive and negative aspects to these online reviews that had emerged over the years. Since their implementation in 1999 to now, large

amounts of data have been collected that show both sides of the concept (Shahare, 2022). The major known benefit of online reviews is that if many satisfied customers left positive Google reviews on a business's website, it would serve as free advertising for the company by placing it in the highest-rated section and appealing to a broader audience. (M., 2025). This draws attention back to the previous point about how these reviews serve as a mutualistic relationship between the two entities at hand, meaning that both the customers and businesses are benefiting. The businesses are benefiting from the higher established credibility and enhanced marketing with limited spending on advertising from their end and the customers benefit because they will know what to expect prior to utilizing a business's products or services for the first time. Along with this, there are a number of other benefits that can occur from the use of online reviews.

Not all the cases from the history of online reviews showed that the positive reviews aided businesses. There were also numerous cases in which the negative, online reviews had broken businesses and crippled the future success of an established entity. One of the major cons that business owners had revealed throughout the years was that just "one negative review of a product or business can skew a potential customer's view of them" (M., 2025). Even if there was one customer who had a subpar experience at a business and left a negative review while the others had positive reviews, a new customer could be fearful to try the place because of the feedback left. In cases such as these, the reviews can serve as negative marketing and sway away potential customers to the business. Through these negative reviews, the businesses have found a way to come out of the trap of a poor reputation by accepting the negative feedback as advice on how to improve their business, and "providing details of how the manager successfully resolved them" (M., 2025) to ensure the same issues do not repeat again. By doing this, businesses have been able to establish trust with their customers and show how the company has fixed the isolated incidents of poor service that have occurred.

The current fate of business reviews are further being studied through sentiment analysis since there is now 25 years of data and reviews that have been collected. Many are rushing to study the effectiveness of reviews through the using machine learning. The established concensus is that the online review strategies are essential for businesses since statistics that show "35% of customers are looking only at companies with four-star ratings or higher" (Podolsky, 2024) and "70% of consumers use rating filters when looking to discover local businesses" (Podolsky, 2024).

These statistics show how reviews have an all time high importance. Through various studies that are emerging, the future of reviews will be determined and tips can be given to businesses on how to use reviews in their favor and strategize accordingly. There are many revolutions to arise in the industry of reviews from all the studies being conducted and many features will be added to further benefit businesses (Podolsky, 2024).

A Harvard business review explains the underlying process behind sculpting customer experiences and their long-lasting effects on revenue. The study highlights that the revelations and transformations that emerge from this data can be alarming. It discusses CEOs who choose to value customer experience management (CEM) over customer relationship management (CRM). For example, CISCO's CEO, John Chambers, prefers to wait for customer reactions before selecting key technologies rather than focusing on competing with others in the market. This approach ensures that they have a market to sell to, as they align their offerings with customer preferences (Meyer and Schwager, 2007). Almost every sector in the

business industry has recognized the value of online customer reviews. A study examined the relationship between customer sentiments in movie reviews and future sales, considering various factors that might influence the perceived helpfulness of reviews. These factors included review length, sentiment polarity, the number of responses received, review subjectivity, and the average rating of all reviews for the movie. (Archak et al., 2007). Another thorough study on online customer reviews of coffee shops in China found that local or independent coffee shops are more sensitive to price fairness than franchise shops, which have stable prices and reputations. This conclusion was reached using a Chi-square difference test on business types. (Tao and Kim, 2022). Another major sector that online reviews have significantly impacted is tourism, where travelers heavily rely on reviews to plan both local and international trips. Guerreiro and Rita (2020) conducted research on predicting customer sentiment using text mining on Yelp restaurant reviews and their financial influence on businesses. A study on the Marriott Hotel in Beijing took an unusual approach to analyzing online negative reviews and complaints while also examining how managers of luxury hotels handled these complaints, whether they followed service recovery methods within the Hospitality industry. This study emphasized the ease of expressing negative reviews on TripAdvisor (the largest hotel review site) and revealed the underlying factors behind customer complaints. The researchers identified the methods such as compensation, which highly influenced customer sentiment and helped prevent negative reviews (Chen and Tabari, 2017). After summarizing findings from different online customer review analyses, Dahiya et al. (2021) explored the layers beneath the current process of mining customer reviews. They demonstrated the numerous methods used and their results, highlighting the influence of star ratings and the importance of quality over quantity in reviews. To determine the effect of customer reviews on a customer's decision to buy from a business, it was found that only 5% of people did not consider reviews important. This finding underscores the value and significance of customer reviews for any customer-centric business in the modern era.

1.3 Methodologies in Literature

As noted before, traditionally, businesses used to rely on word-of-mouth (WOM) to influence product sales. WOM occurred through interpersonal conversations, but with the advent of online review systems, consumers can now share their opinions freely and widely, impacting potential buyers' decisions significantly (Shen, 2008).

A study by Pinto et al. (2024) employs the UCI ML Drug Review dataset, which contains a significant volume of user-generated content related to drug experiences. The study emphasizes the importance of EDA in understanding the distribution and characteristics of the data. The paper also utilizes two prominent sentiment analysis tools: TextBlob and VADER. These tools are employed to categorize the emotional tone of patient reviews. For our purposes, this process can be used to quantify the sentiments from reviews, which can be useful to the businesses and stakeholders.

Reddy et al. (2024) highlight the use of both traditional machine learning models like Naive Bayes and Support Vector Machine (SVM) along with some advanced techniques like deep learning architecture (e.g., Recurrent Neural Networks and Transformer models like BERT). This comparison is crucial for understanding the evolution of sentiment analysis methodologies. The study also identifies several challenges with sentiment analysis, such as ensuring

data quality, handling language variations, and addressing privacy concerns. The study notes that sentiment analysis can play a significant role in risk management by identifying potential issues early on.

Kyriakidis and Tsafarakis (2024) proposed an integrated framework that combines multiple analytical techniques to effectively process customer reviews. The technique helps in identifying specific aspects of products or services mentioned in reviews and determine the sentiment associated with that aspect. This allows businesses to understand customer opinions on particular features or services. This method incorporates fuzzy logic to handle the uncertainty and imprecision, which is often present in customer reviews. The thing which prompted us to incorporate this paper in our project is the fact that this framework is designed to process both textual and numerical data derived from online reviews. This will help us get a more holistic view of customer feedback.

Guerreiro and Rita (2020) used a lexicon-based approach to label words in a lexicon as positive or negative based on their semantic context. They used IBM SPSS Modeler Text Analytics as the sentiment analysis tool. This tool automatically structures text into groups of words based on contextual and semantic information. They created a base dictionary, which included part-of-speech codes for each term. This dictionary was expanded with synonyms to classify and label the text more effectively. The authors categorized terms into 17 different sentiment categories, including both positive and negative connotations. They found the CHAID decision tree model to be effective. This method can be thought of as a more advanced approach to the simple Bag-of-Words text classification approach.

Some studies like Pinto et al. (2024) mentioned that we can get more refined insights with enhancements in data pre-processing steps. Also, exploring different types of sentiment analysis tools can give more perspectives to see the customer reviews. One potential limitation that we may face and as noted by Chen and Tabari (2017) Chen, is that there could be some self-bias among customers, and due to that, a customer may provide a highly negative or positive review for a particular business. This self bias can also manifest in the customer population, as all customer reviews may not be real and some can be from competitors. Another significant takeaway is from a paper by Kumar et al. (2024), which is that the selection of machine learning algorithms for sentiment analysis should be determined by the overall result that we want to achieve, which can include the nature of the dataset and various other factors.

1.4 Project Plan

We aim to explore how online reviews can serve as a powerful tool for analyzing business performance, offering valuable insights into customers sentiment, market trends , and competitive positioning. Through the application of text mining, machine learning, and sentiment analysis techniques, businesses can extract meaningful patterns from unstructured data, allowing them to make data-driven decisions.

2 Methods

2.1 Methods: Data Source and Munging process

We have compiled our dataset using the Google Local Data review and metadata files for each of the 50 states across the United States provided by UCSD's Tianyang Zhang and Jiacheng Li (Yan et al. (2023), Li et al. (2022)). We first started by creating a pipeline to extract both JSON files per state and store it into Google Cloud Platform (GCP) for further transformation. We decided to use GCP for storage efficiency and accessibility. Once we had all 100 files in a GCP bucket, we proceeded to create our main pipeline which would take the JSON files directly from GCP and create a single dataframe for each state (combining the metadata and review files) by performing an inner join on the unique *gmap_id* column, merging all the data frames together into one primary dataframe and then saving it in the form of a parquet file back to GCP.

Alabama	reviews (8,967,499 reviews)	metadata (74,967 businesses)
Alaska	reviews (1,051,246 reviews)	metadata (12,774 businesses)
Arizona	reviews (18,375,050 reviews)	metadata (108,579 businesses)
Arkansas	reviews (5,106,056 reviews)	metadata (47,246 businesses)
California	reviews (70,529,977 reviews)	metadata (515,961 businesses)
Colorado	reviews (15,681,222 reviews)	metadata (106,829 businesses)
Connecticut	reviews (5,181,800 reviews)	metadata (49,200 businesses)
Delaware	reviews (1,885,948 reviews)	metadata (14,706 businesses)
District of Columbia	reviews (1,894,317 reviews)	metadata (11,060 businesses)
Florida	reviews (61,803,524 reviews)	metadata (378,020 businesses)
Georgia	reviews (24,060,125 reviews)	metadata (166,381 businesses)
Hawaii	reviews (3,111,531 reviews)	metadata (21,507 businesses)
Idaho	reviews (3,892,636 reviews)	metadata (33,214 businesses)
Illinois	reviews (23,096,838 reviews)	metadata (179,205 businesses)
Indiana	reviews (12,865,167 reviews)	metadata (100,391 businesses)
Iowa	reviews (4,838,887 reviews)	metadata (47,794 businesses)
Kansas	reviews (5,546,880 reviews)	metadata (46,286 businesses)
Kentucky	reviews (7,654,993 reviews)	metadata (63,193 businesses)
Louisiana	reviews (7,536,078 reviews)	metadata (63,315 businesses)
Maine	reviews (2,214,773 reviews)	metadata (24,853 businesses)
Maryland	reviews (10,728,483 reviews)	metadata (78,144 businesses)
Massachusetts	reviews (10,447,007 reviews)	metadata (92,520 businesses)

Figure 1: Snippet of the JSON files.

There were a few other modifications we made to the dataframes through our pipeline before merging them, like we dropped the $MISC$ column due to its inconsistent datatype throughout different states. We also decided to drop duplicate *gmap_id* values to ensure data accuracy and rename the column *name* to *business_name* in reviews dataframe and *customer_name* in metadata to avoid data redundancy and clarity. Other than that, we also added a column *state* while merging the data frames for each state for state identification and further regional classification.
In order to shorten the runtime of this code and maintain the connection to GCP cloud, we created smaller subsets including 5-8 states each and saved the parquet files for each small

subset in GCP, then performed the final step of merging all the parquet files into a single dataframe. Once we had our final dataset, we started with the data cleaning.

For the data cleaning aspect of the exploratory data analysis, our team has cleaned up various aspects of the dataset by dropping columns, adding columns, renaming columns, etc., in order to make analysis easier and to be able to provide higher quality insights from the data we have collected. We first started with renaming the columns, like *text* column to *reviews*.

The next step of the process was to remove any unnecessary columns that the dataset had contained. Some of these columns included: *relative_results*, *pics*, *resp*, *price*, and *description*. We have removed these columns because we have noticed that these categories had a lot of null values and would not be able to provide any actionable insights into our study. Although we could have answered some questions about business performance using these columns, the null values would have skewed the results and shown false trends due to the vast amount of missing data in these columns. After dropping these columns, we were able to drop all of the null values from the remaining columns that we had decided to keep for our study. With doing these steps in this order, we were able to prevent mass amounts of data from being removed from the study.

The next few columns that have dropped included: *url*, *address*, and *user_id*. With these columns, we would not be able to complete any further analysis and these are just characteristics that uniquely identify each of the businesses. Since these are considered as primary keys and identifiers, they will not provide any meaningful visualization when these are plotted against any of the other variables in the study.

After we have finalized all of the columns we were going to drop and deemed them unnecessary to the study, we have added a region column. Using a map of the United States, we have categorized each of the states into different geographical locations based on their relative location considering all states. All states have been added to one of the six regions of the United States: Midwest, Northeast, Southeast, Southwest, West, and the Midwest. We have added this column since we will be able to use it to perform further analysis on the geographical region and its effect on business performance levels.

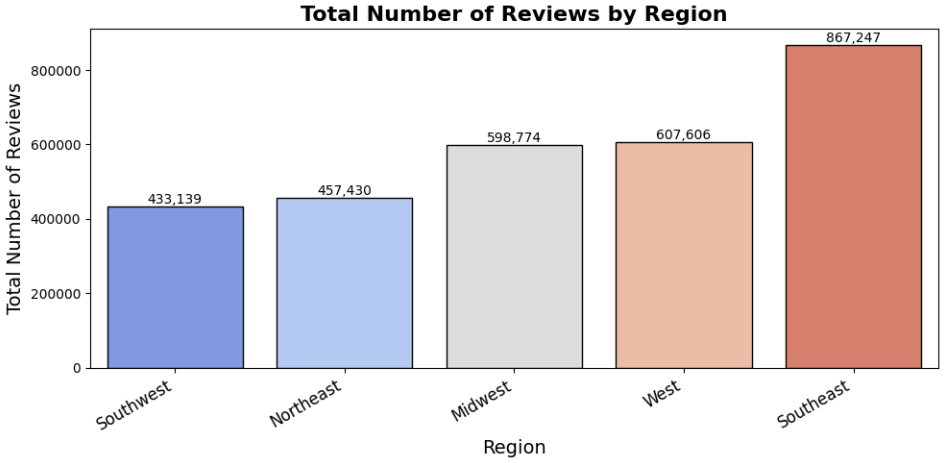


Figure 2: Distribution of Review by Region.

After adding the region column to our dataset, we graphed the number of reviews from each of the regions in order to see the distribution of reviews across the United States. It can be seen that the Southeast region had the highest number of reviews and the Southwest region had the lowest number of reported reviews in the dataset.
The next aspect of the data cleaning process included creating a *standard_category* column. Our dataset already had a *category* column, but there were way too many categories in order to do any meaningful analysis. To create the *standard_category* column, our team utilized common key words and strings to categorize each business into a specific *standard_category*. For example, if there were key words such as, "restaurant", "food", "deli", "takeout", "sandwich", etc., we have combined all of these into a combined "Restaurant" *standard_category*. Similarly, we did this same procedure for other industries as well and created *standard_category* such as "Healthcare", "Automotive", "Transportation", etc. This way, we would only be dealing with fewer, general categories rather than numerous smaller categories which all represent the same type of industry.
Using the "from_unixtime" functionality of pyspark, we were able to process the time column of the dataset. With the "from_unixtime" function, we extracted the week, month, and year from the times listed in the dataset to create a more specific timestamp that will allow us to perform analysis of time variables effects on business performance. After performing this step, we realized that the year aspect was incorrect because it was in milliseconds; therefore, we have divided these values by 1000 to convert to seconds and get an accurate timestamp.

2.2 Methods: Overview of Review Trends Over Time

Knowing how review volume varies over time is important for detecting larger patterns in consumer behavior. Our analysis centers on the temporal distribution of online reviews from 2015 to 2021. While the dataset includes reviews from previous years, we restricted the main visualization to this six-year period because entries prior to 2015 account for less than 0.1% of the data and provide limited interpretive insight.

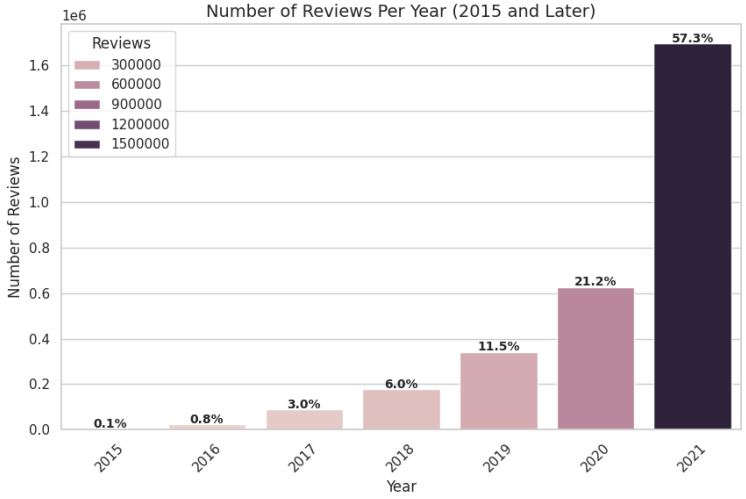


Figure 3: Total Count and Percentage of Online Reviews from 2015 to 2021.

Since 2015, the volume of reviews uploaded each year can be observed to increase consistently. This trend is reaffirmed in 2018 and then continues throughout 2019, as consumers increasingly seek out online sources as a means of influence for their interactions with companies. The highest jump is witnessed between 2020 and 2021.

Figure 4: Fluctuations in Review Counts between January 2019 and September 2021.

In Figure 4, we can see a stable monthly trend throughout 2019, with relatively small changes. However, at the beginning of 2020, the number of reviews showed a steady upward trajectory. This trend accelerated rapidly in early 2021, reaching the highest peak around April. This peak, with nearly half a million reviews, confirms the presence of specific influences or events that significantly altered consumer engagement during that period. Following the peak in April 2021, monthly review counts declined, returning to baseline levels.

2.3 Methods: Understanding Length of Reviews and Affect on Rating

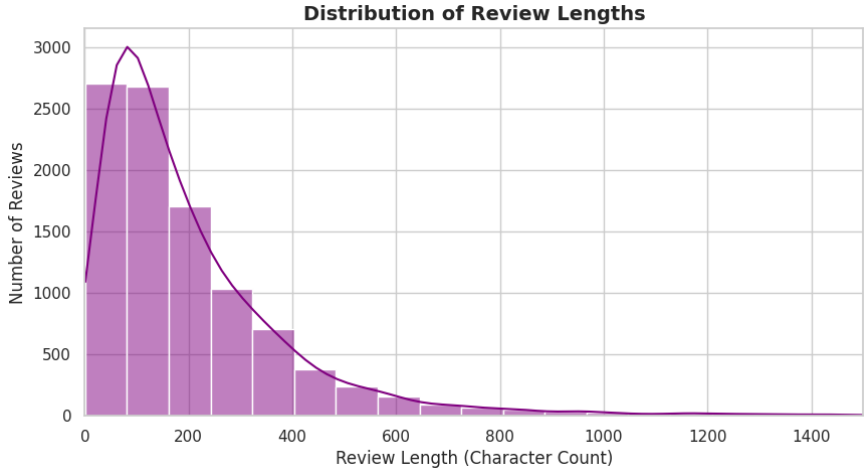


Figure 5: Distribution of Review Lengths (Character Count).

From Figure 5, it can be seen that the distribution is right skewed with a peak at around a character count of 100. This means that a majority of the reviews fall under this length and the outliers, which are causing the skew, fall under a 800+ character count. From this, we know that customers typically leave a short review, with an exception of a few outliers which leave character counts of more than 1000.

2.4 Methods: Analyzing Business' Open/Close Rates

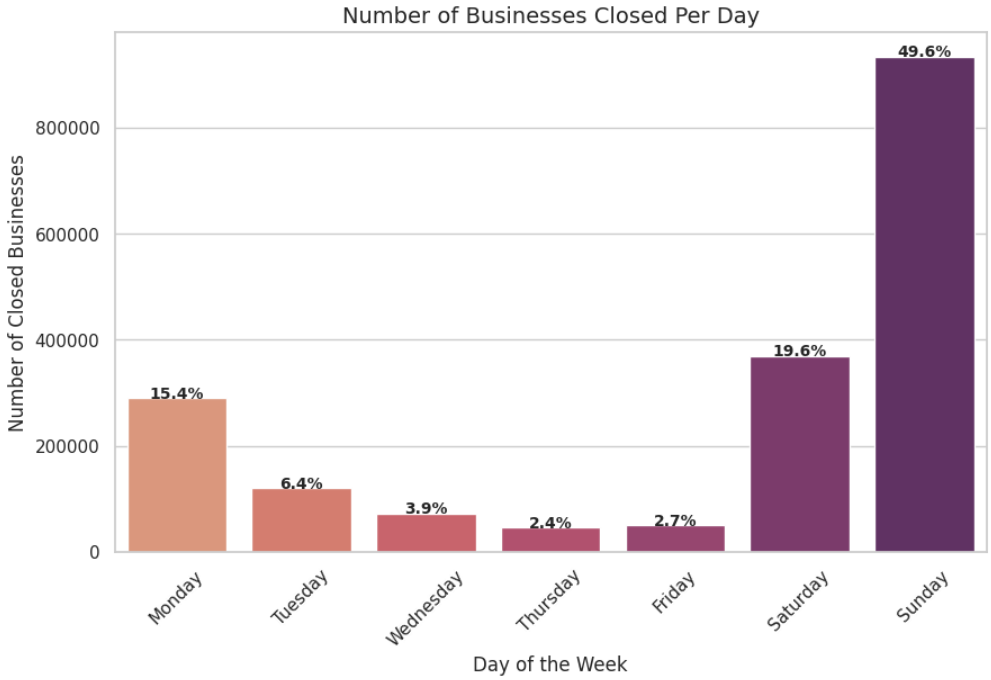


Figure 6: Distribution of Businesses Closed over each Day of the Week.

From Figure 6, we can conclude that almost half (49.6%) of businesses are closed on Sunday. Saturday (19.6%) and Monday (15.4%) also show relatively high closure rates. Midweek days (Tuesday - Friday) have the lowest closure rates, suggesting most businesses are active during weekdays. For a business, it is important to understand the most active business days, as usually, the customers are also active on those days as well. But it also depends on the business industry or category. Some businesses like Hotels tend to be more active on weekends because more people tend to visit them over the weekend. So we have to analyze the category specific closure rates. We also have to be careful about not just showing the number of businesses in each category, as it will create bias towards categories with less businesses. Therefore, we can find the percentage of businesses closed in each category.

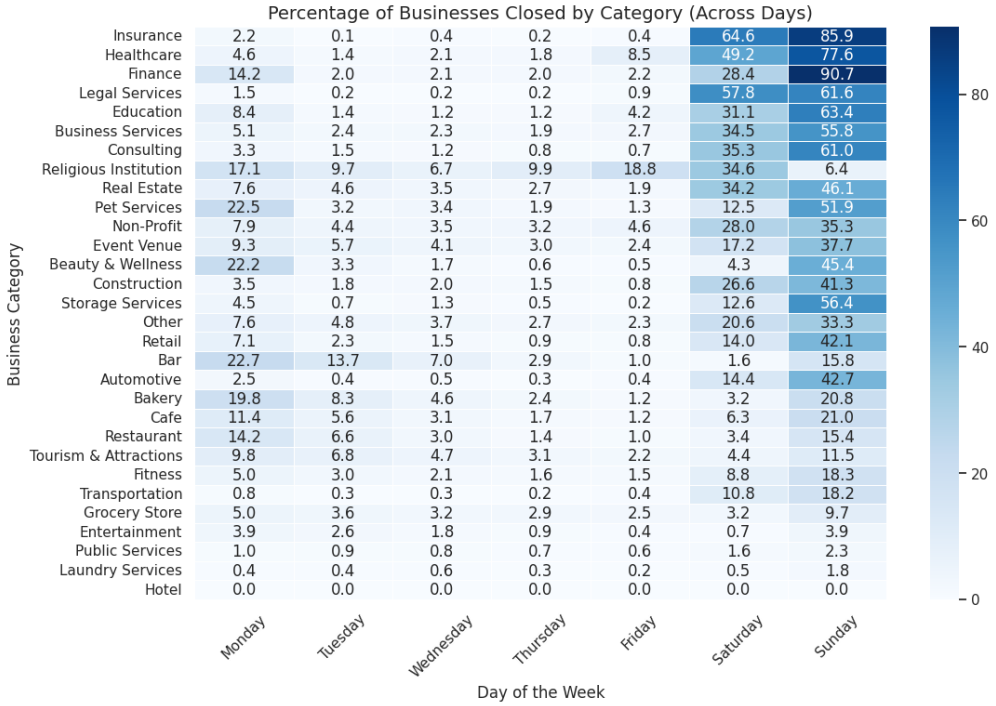


Figure 7: Percentage of Businesses Closed per Category across Weekdays.

As we suspected, industries like Entertainment and Hotel tend to be the most active on weekends, while Retail and Restaurant tend to be closed over the weekend. Another thing to note is that religious institutions like Church are usually open on Sunday, and closed on other days of the week. This pattern is very obvious, as most people visit the Church on Sundays. Overall, this can help new businesses to thrive, as it shows the activity of businesses, and therefore its customers over the week, using which they can determine their business hours.

2.5 Methods: Regional Trends Analysis

2.5.1 Business Density and Hotspots

This visualization Figure 8, effectively shows where businesses are concentrated across the United States and helps identify urban hotspots around metropolitan cities like New York, Chicago, Los Angeles etc. The intensity level is described using darker color, which indicate higher intensity, which helps compare density across urban centers, while the grey marks identify businesses in less populated areas to show the clear distinction in distribution of businesses in midwest region, in comparison to the west and east coast. Other than that, we can also observe that business density follows major highways, for example, Texas (Dallas-Houston) and California (SF-LA).

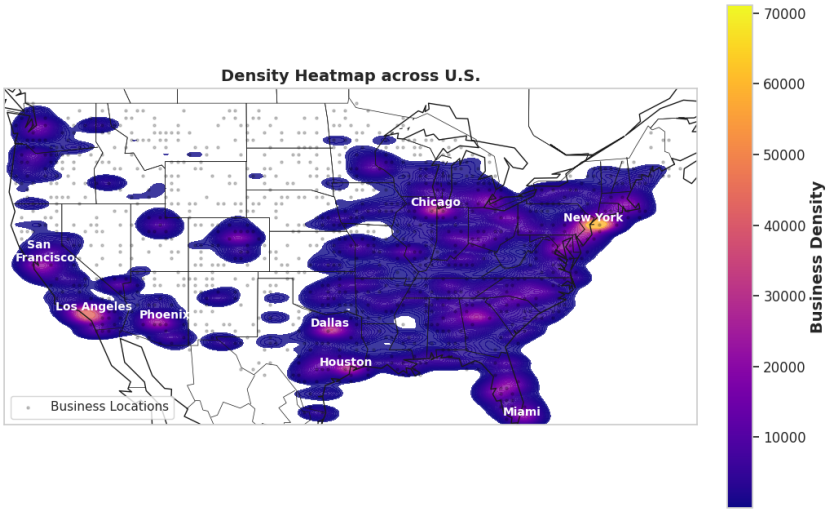


Figure 8: Density Plot highlighting Business Hotspots across U.S.

2.5.2 Average Ratings by State

This map visualizes the average business rating for each U.S. state, with color coding to show rating variations. Northern states tend to have the highest ratings while the southern states like Texas, Arizona, Georgia etc fall below the median rating, which could be due to higher number of businesses leading to mixed reviews. Overall, most states have high average ratings pointing towards the positive sentiment bias in our dataset. But there is also a crucial regional difference based on the customer ratings can be observed.

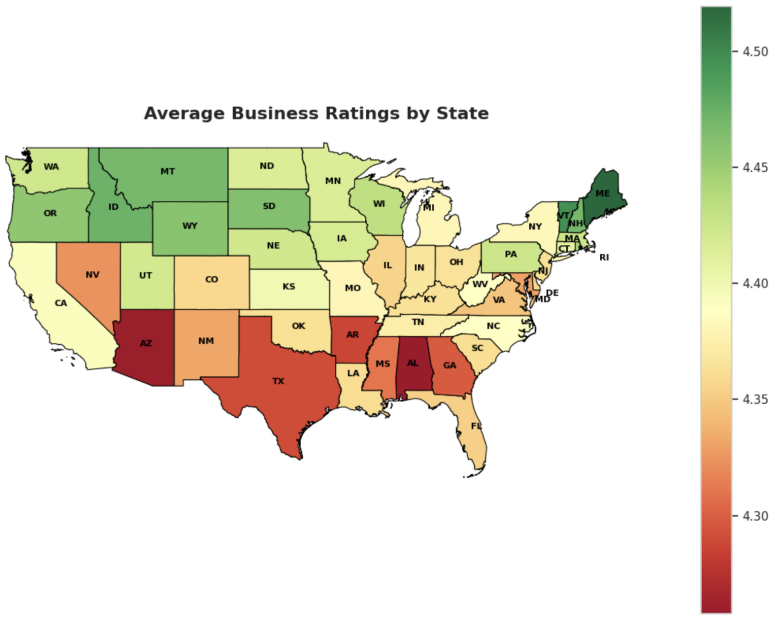


Figure 9: US Map showing trend in Average Ratings.

2.6 Methods: Correlation between Average Rating and Number of reviews

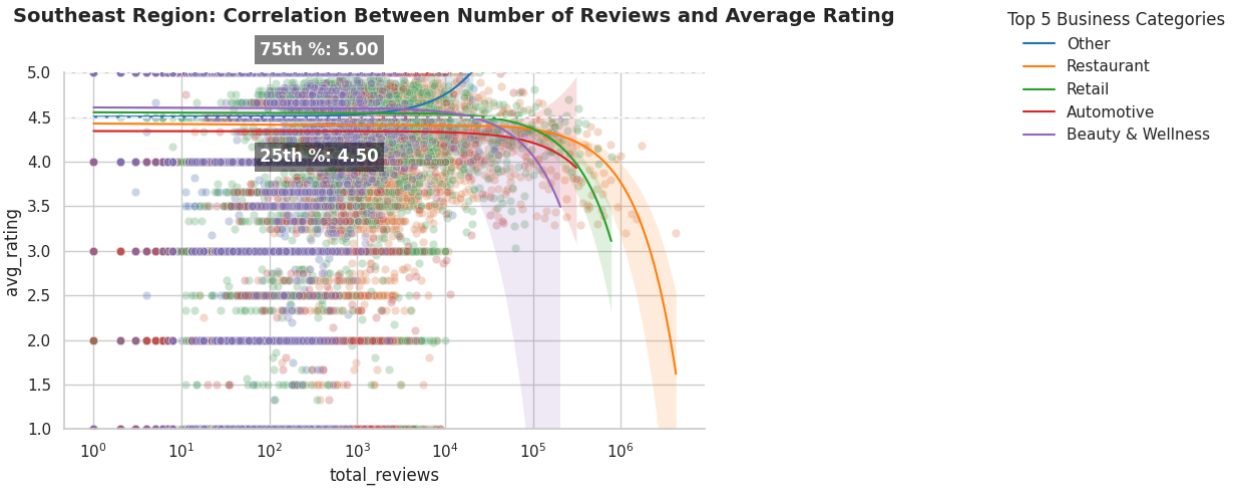


Figure 10: Scatter plot with trend lines showing the correlation between Average ratings and Number of reviews for a business.

Through this analysis of visualizing the trend between Average rating and Number of reviews for Southeast region (region with the most number of reviews) we were able to understand that businesses with more reviews tends to have slightly lower ratings. The scatter plot and trend lines show that businesses with fewer reviews tend to have a wider spread of ratings, however as number of reviews increases, average ratings tend to stabilize around 4.0-4.5 but at very high review counts, they slightly decrease, as shown by the trend lines. This signifies that larger or more popular businesses with more reviews don't necessarily mean a higher rating. We also want to note that this pattern was observed in all the regions.

2.7 Methods: Sentiment Analysis

For sentiment analysis, we started with rule-based models. For rule-based models, there are predefined rules and sentiment lexicons. Each word in a review is looked up in a sentiment lexicon to retrieve a polarity score or label. The review's overall sentiment can be computed by summing or averaging these scores. We started with two specific rule-based models: VADER and TextBlob.

2.7.1 Methods: VADER

Vader (Valence Aware Dictionary and Sentiment Reasoner) is a pretrained sentiment analysis model which specializes in Social Media text as it was initially developed to classify tweets sentiment. It starts with analyzing the polarity of words and assigning a sentiment score to each word. It combines the positive, negative and neutral proportion of the text to give an aggregate score called "compound" which is computed as a normalized value between

-1 (very negative) to +1 (very positive). As our main goal is to classify customer reviews, VADER takes into account emoticons, slang and capitalization.
We decided to classify the sentiment based on:

- Compound score > 0.05: Positive sentiment
- Compound score < -0.05: Negative sentiment
- Compound score between -0.05 and 0.05: Neutral sentiment

2.7.2 Methods: TextBlob

TextBlob is also used for processing textual data. It provides an easy-to-use API for some common NLP tasks, which includes sentiment analysis too. TextBlob gives two outputs; one is "polarity", which measures the sentiment of the text. The range of polarity is from -1.0 (negative) to +1.0 (positive). The second one is "subjectivity", which provides confidence in the expressed sentiment. In simple terms, confidence measures how opinionated vs. factual the text is. The range of subjectivity is from 0 (objective) to 1 (subjective).
Initially, we attempted to combine polarity and subjectivity into a single score to capture both tone and confidence. We used the following formula for that:

$$\texttt{combined\_score} = (\texttt{polarity} \times \texttt{subjectivity} + 1) \times 2$$

But because of that, most reviews were labelled as "neutral" based on the labels we created. A likely reason for that might be due to subjectivity scores weighing down meaningful polarity. Therefore, we simplified our approach to rely on polarity alone, which led to better distribution across all sentiment classes.

- Polarity score > 0.05: Positive sentiment
- Polarity score < -0.05: Negative sentiment
- Polarity score between -0.05 and 0.05: Neutral sentiment

2.8 Methods: Sentiment Annotation Experiment

2.8.1 Introduction

In this experiment, we aimed to explore the performance of sentiment analysis tools like TextBlob and VADER, and compare that to the reviews labelled by our group based on the sentiment. We used a diverse set of 30 reviews from our dataset. The goal was to understand how these lexicon-based models align or diverge from human interpretation, and to identify common areas of disagreements.

2.8.2 Review Selection Methodology

Our dataset contains almost 3 million rows. It was impossible for our group to label that many ratings, and after further discussion, we came to the conclusion that from labelling a few reviews and making a comparative analysis, we can gain some key insights on the performance of TextBlob, VADER, and other models that we will use in the future.
To ensure a balanced and somewhat representative sample from such a large dataset, we implemented a stratified sampling strategy using PySpark.

Key Sampling Criteria:

- Sentiment Polarity (VADER-based): Positive, Negative, Neutral
- Review Length: Short (< 100 chars), Medium (100–300), Long (> 300)
- Review category: Picked from top 4 most frequent categories (Retail, Restaurant, Beauty & Wellness, Automotive)

We used VADER's sentiment scores to assign the initial sentiments, and reviews were bucketed by length. From each category, 1 review was samples for every (`sentiment label * length bucket`) combination. This code logic gave us a diverse set of 30 reviews.

2.8.3 Human Annotation Design

Each of the 30 reviews were rated by each member of our group (resulting in 4 labels). We created the ground rules for annotation to maintain consistency and reduce subjectivity. We defined 19 rules for labelling reviews. These rules covered:

- Tone-based interpretation
- Handling emojis, intensifiers, sarcasm, and factual statements
- Protocol for mixed-sentiment or ambiguous cases
- Weighting the positive/negative elements in long reviews

2.8.4 Limitations and Potential Biases

While this exercise gave us some quality insights, several limitations must be addressed:

- Small sample size: We labelled only 30 reviews from almost 3 million reviews.
- Selection bias: Sampling still relied on VADER for stratification.
- Group annotation subjectivity: Despite the ground rules we made, we acknowledge that sentiment interpretation varies.

From this experiment, we now have a set of reviews labelled by our group. This can be considered as a "baseline" or "gold standard", and can further be used to evaluate the performance of more advanced sentiment analysis models that we will be using.

Our experiment results somewhat relate well with the findings of other studies. Some studies show that both VADER and TextBlob perform well over social media text, with VADER slightly outperforming TextBlob. But in the context of reviews, TextBlob performs better than VADER due to its larger lexicon (Alemán Viteri, 2021).
Due to the inconsistencies between TextBlob and VADER, we decided to try a transformer-based, zero-shot classification approach using facebook/bart-large-mnli.

2.9 Zero Shot Text Classifier - Using Pre- trained BART Model

Zero Shot Classifier uses pre-trained language models that are capable of understanding context deeply. According to Yin et al. (2019), zero-shot text classification is particularly useful for settings with no task-specific training data, allowing generalization across domains and tasks.This approach helps eliminate the shortcoming of unlabelled data by providing an engineering technique that allows a model to perform tasks without specific training.
A widely used and highly recognized zero-shot model is BART (Bidirectional and Auto-Regressive Transformers) model, which was developed by Facebook (now Meta AI). Since BART is a transformer-based model, it can perform a wide array of tasks, like summarization, translation, generation, etc. BART uses both bidirectional encoder (like Google's BERT) and autoregressive or left-to-right decoder (like GPT) (Lewis et al., 2019).
The Towards AI article by Giancaterino (2023) demonstrates that facebook/bart-large-mnli handles domain-specific classification effectively and consistently performs well across unseen tasks, making it a strong candidate for sentiment classification without training.
We tested BART Large MNLI, where BART is trained on the MultiNLI (Multi-Genre Natural Language Inference) which contains about 433k crowd-sourced collection of sentence pairs, and is one of the largest corpora available for NLI (Natural Language Inference) Williams et al. (2018). This helped us with the zero-shot approach as the model was already pre-trained and did not require any more training or fine-tuning.

2.10 Labelling Reviews using ChatGPT 4o (API)

We explored ChatGPT 4o API to label a subset of reviews. We selected another sample of 10k reviews from our larger dataset using stratified sampling to ensure balanced coverage across:

- Review Lengths
- Star ratings
- Business Categories

NOTE: We went with a stratified sampling approach because it gives a wide variety of samples from our large dataset, and training or testing different models on this would be easier for us.
Each review was passed into the API with the following instruction:
Label the following customer review as positive, negative, or neutral based on its tone and meaning.
The output from the API was parsed and mapped to numerical classes:

- Positive: 2
- Neutral: 1
- Negative: 0

2.11 Methods: Customer Engagement Forecasting for Businesses

The goal was to predict whether a business's engagement in 2021 increased, decreased, or stayed stable based on review and operational trends from previous years (2018–2020). For this, we chose businesses which got reviews in all four years (2018-2021). This implied that there is some customer engagement across all four years for that business. We found 5608 businesses which received reviews across all four years, and there were around 658k reviews associated with those businesses in our dataset.
One problem was to label all those reviews using ChatGPT API. It would have been very expensive for us, both in terms of cost and resources.
To solve this problem, we decided to fine-tune a DistilBERT model on 10k stratified review labels generated by ChatGPT.

Fine-tuning process:

- Base model: `distilbert-base-uncased` (from HuggingFace Transformers)
- Tokenized using `DistilBertTokenizerFast`
- Training configuration:
  - Batch size: 16
  - Epochs: 3
  - Optimizer: AdamW
  - Learning rate: 2e-5
- Evaluation metric: Macro F1-score

We chose Macro F1-score as the evaluation metric because it gives us a more balanced and representative score, unlike Weighted F1-score, which is highly susceptible to class imbalance. The model was trained using Trainer from HuggingFace, with validation monitoring after each epoch.
The BERT model was applied to all reviews in the dataset of businesses which received reviews in all four years (2018-2021). Sentiment predictions were:

- Averaged by business-year
- Integrated into the composite engagement score per business per year

We calculated Composite Engagement Score by normalizing average yearly ratings, average yearly sentiment, review count, and average number of open days, and taking a sum of those four variables.
To check whether there is a class imbalance in composite score or not, we checked its distribution:

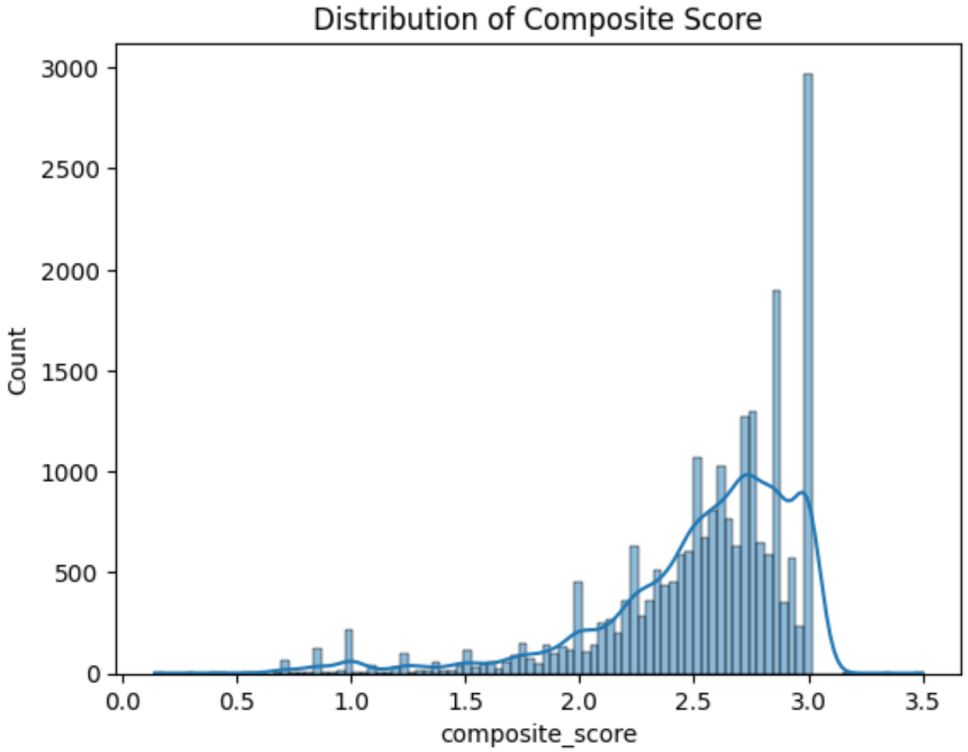


Figure 11: Distribution of Composite Score.

There is a clear imbalance, as there are far more businesses with higher composite scores than lower. We will be accounting for this while creating the model.

To capture engagement trends, we created delta features:

- `delta_18_19 = composite_score_2019 - composite_score_2018`
- `delta_19_20 = composite_score_2020 - composite_score_2019`

These deltas represent the trajectory of a business's engagement, providing the model with temporal patterns.
We then calculated percent change in engagement between 2020 and 2021:

```
pct_change_20_21 = ((score_2021 - score_2020) / score_2020) * 100
```

Labels were assigned as follows:

Label	Description	Rule
0	Engagement Decreased	$\leq -20\%$
1	Engagement Stable	$> -20\%$ and $< 20\%$
2	Engagement Increased	$\geq 20\%$

Table 1: Label definitions for engagement change based on percentage rules.

Model Selection:
We selected **XGBoost** for its effectiveness on tabular data and its flexibility for multi-class problems. We selected a few model parameters based on our dataset, and they are as follows:

Parameter	Value	Reason
objective	multi:softmax	Predicts one of 3 discrete classes
num_class	3	Our target classes are 0, 1, 2
eval_metric	mlogloss	Suitable for multi-class classification

Table 2: Key parameters used in model training.

Since there was a class imbalance as we observed, we handled it using `sample_weight`, computed via `compute_sample_weight(class_weight='balanced')`.

The model was trained on 99% of the training data (after a 20% test and 20% validation split) using `StratifiedShuffleSplit`.
We used 99% instead of 100% because:

- `train_size=1.0` is not allowed in stratified splits, as it requires at least one row out of the shuffle
- 99% provides nearly all data while ensuring balanced sampling

2.12 Methods: K-Means Custering Based on Sentiment, Rating and Volume

Using the unsupervised machine learning algorithm K-means clustering, we decided to group similar businesses together based on their average rating, number of reviews to account for popularity and sentiment label using the BERTdistill model previously mentioned. Due to a huge volume of data, we decided to focus on a single market, identified as a container of significant urban hotspots using the Geopandas density Heatmap - **California**.

2.13 Methods: Business Recommendation Tool - KNN based

We built a Business Recommendation Tool using KNN to find similar businesses based on review patterns, sentiment, business category (type of business) and the location of a particular business. This is a versatile tool with results useful to both customers that want to find similar businesses based on their performance and businesses wanting to understand their competitor's performance and sentiments. The tool utilizes cosine similarity across TF-IDF vectors of reviews and metadata (such as average rating, sentiment scores, category, and active days) and is restricted by a 25km geographic radius using the Haversine distance formula. Other than that it also uses TF-IDF to find the significant common words from reviews to then present in the form of a word cloud while also providing a barplot comparing the average rating and proportion of Negative, Neutral and Positive reviews of each business including the chosen business.

3 Results

3.1 Results: Sentiment Analysis

3.1.1 VADER

Given below is the sentiment distribution of reviews for VADER:

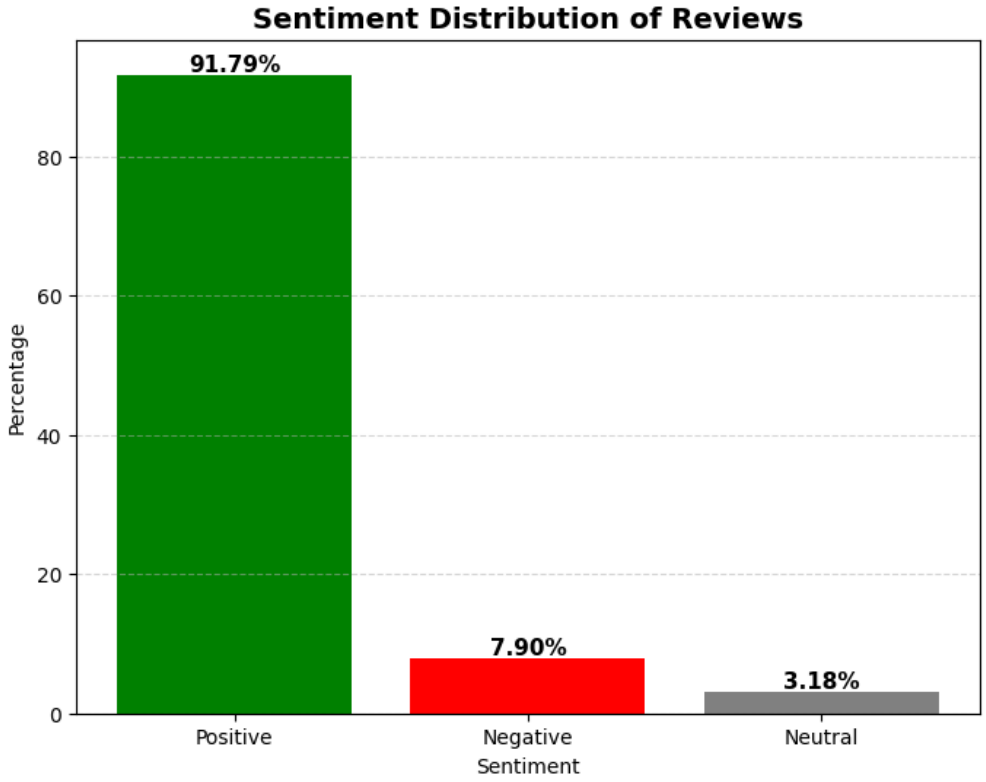


Figure 12: Sentiment Label Percentage for VADER.

The bar chart shows that a vast majority of reviews (92%) were classified as positive and less than 8% were negative while neutral sentiment accounted for just 3%.

3.1.2 TextBlob

Given below is the sentiment distribution of reviews for TextBlob:

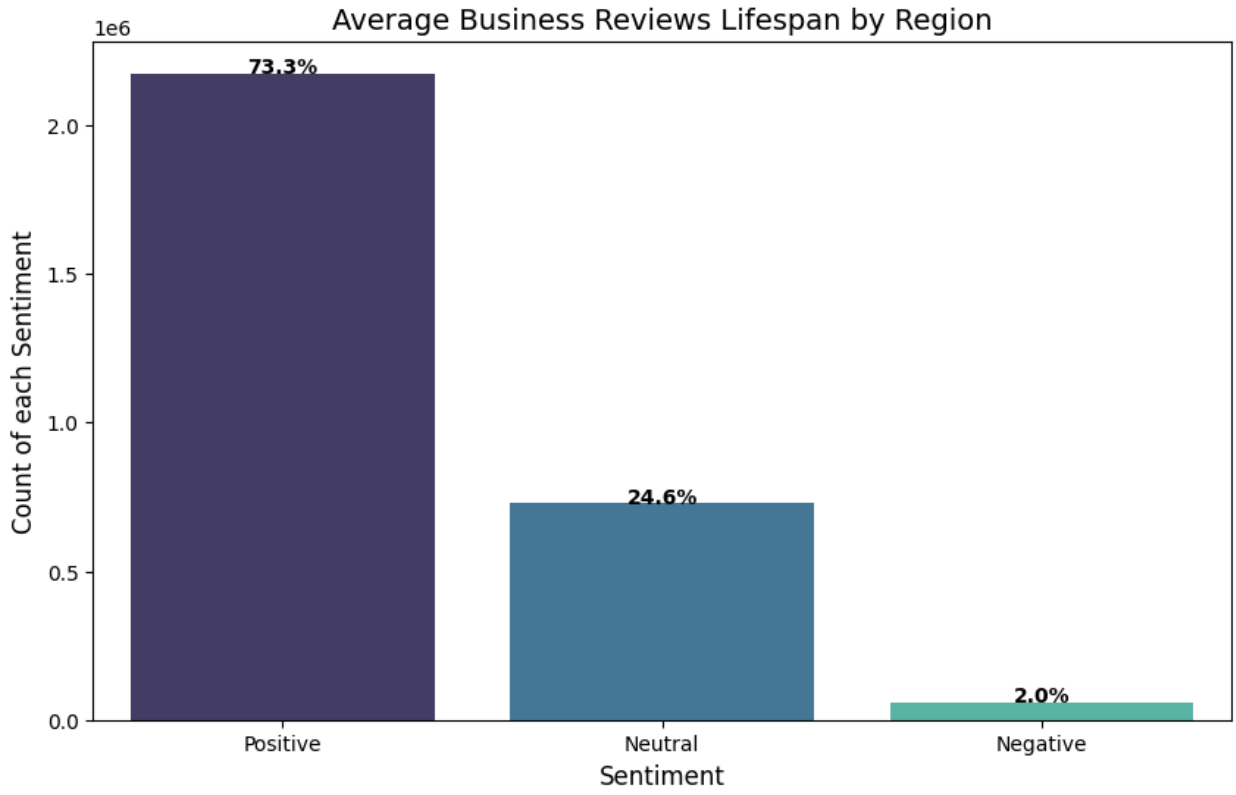


Figure 13: Sentiment Label Count for TexbBlob.

From Figure 13, we can see that most reviews fell into the "Positive" category, which indicates overall positive customer sentiment.
Since we did not have any baseline based on which we can check these models and the models we utilized further, our group decided to do sentiment annotation for a few reviews.

3.2 Sentiment Annotation Experiment

Each review got 7 labels in total:

- VADER label
- TextBlob label
- 4 individual human labels
- An implicit final group consensus (majority agreement or tie-breaker)

The graph below shows the distribution of annotations done by the group members.

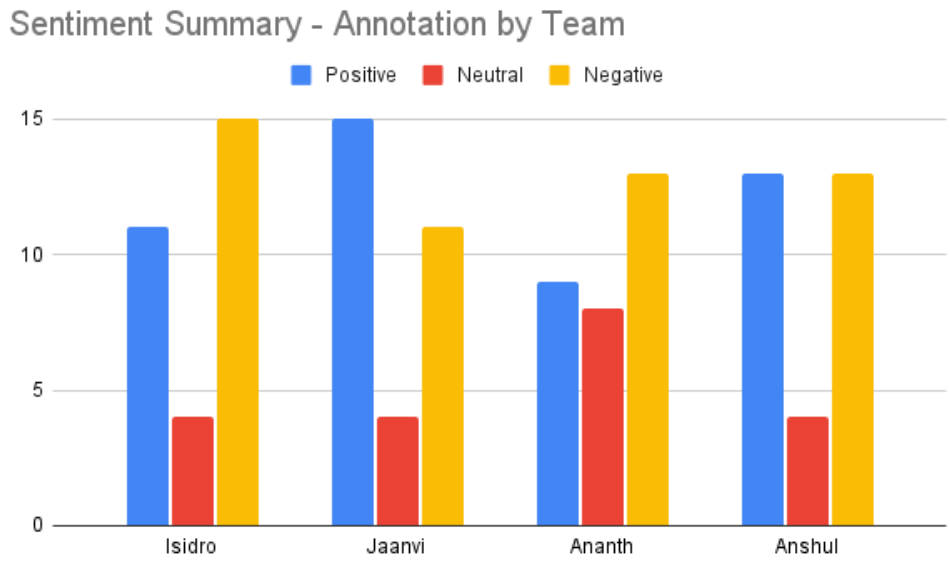


Figure 14: Sentiment Summary - Annotation by Team.

Label Agreement:

- **Full agreement (all 4 agreed):** 20
- **Majority agreement (3 out of 4):** 4
- **Split (2 vs 2 or unclear):** 6

We then as a group reviewed the labels, and reached the agreement on the review labels which were ambiguous.
By doing this, we were able to create a gold standard (or true baseline) against which we can evaluate the performance of a variety of sentiment classification models.
We then compared VADER and TextBlob predictions against our team's final label for each review. Given below are the observations:

(i) Performance comparison on all 30 reviews

	VADER	**TextBlob**
TRUE	18	21
FALSE	12	9
Correct_Pred	**60%**	**70%**

Table 3: Performance of VADER and TextBlob on all 30 reviews

Based on the labels on these 30 reviews by these models, we can see that TextBlob is working better by 10%.

But we wanted to check their performance further, so we compiled 16 reviews where either one or both models have mislabelled, and then calculated the amount of true and false labels for each model.

(ii) Performance comparison on reviews where either VADER mislabelled, or TextBlob mislabelled, or both

	VADER	**TextBlob**
TRUE	5	7
FALSE	11	9
Correct_Pred	**29.41%**	**41.18%**

Table 4: Performance of VADER and TextBlob on a subset

On these reviews as well, the TextBlob is performing better than VADER by 11.77%.

3.2.1 Zero-Shot Text Classifier - Using Pre- trained BART Model

We tested it on our labelled dataset of 30 reviews (baseline), and found that BART's performance is very impressive, compared to VADER and TextBlob.

(i) Performance comparison on all 30 reviews

	BART	**VADER**	**TextBlob**
TRUE	27	18	21
FALSE	3	12	9
Correct_Pred	**90%**	**60%**	**70%**

Table 5: Performance of BART, VADER, and TextBlob on all 30 reviews

(ii) Performance comparison on reviews where either VADER mislabelled, or TextBlob mislabelled, or both

	BART	**VADER**	**TextBlob**
TRUE	15	5	7
FALSE	1	11	9
Correct_Pred	**93.75%**	**29.41%**	**41.18%**

Table 6: Performance of BART, VADER, and TextBlob on a subset

We can clearly see that BART is outperforming both TextBlob and VADER by a huge margin. But, we noticed a pattern that BART was not labelling "Neutral" for any reviews. Therefore, we decided to test it on a larger subset, to see whether BART is labelling the reviews as neutral or not. We tested it on a random sample of 10k rows from our dataset, and noted the distribution of TextBlob, VADER and BART. Given below is the sentiment distribution over 10k random samples:

Figure 15: Sentiment Distribution of TextBlob, VADER, and BART

Our intuition was correct, and even on the larger dataset, BART is not labelling any review as "Neutral". This means that we cannot use BART as a silver-standard for our further models.
NOTE: This also implies that our gold-standard, the set of 30 reviews labelled by our group, contains a lot more positive reviews than negative and neutral reviews.

3.2.2 Labelling Reviews using ChatGPT 4o (API)

Given below is the distribution of ChatGPT on the random 10k stratified reviews:

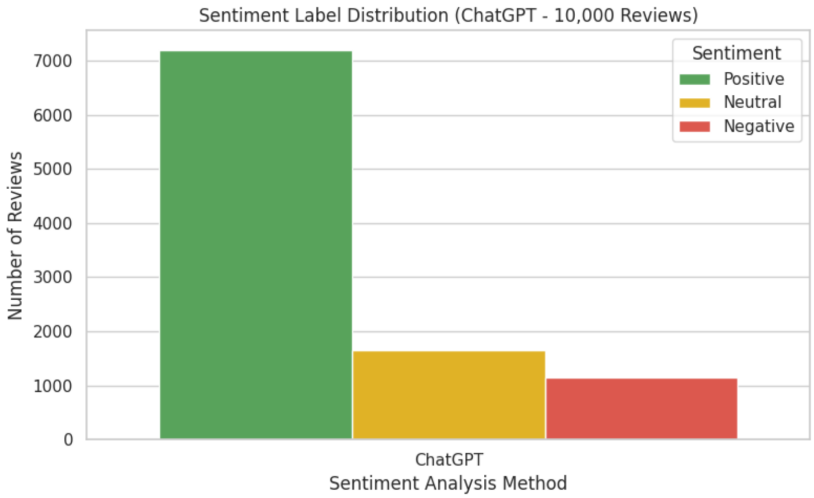


Figure 16: Sentiment Distribution of ChatGPT 4o

We can clearly see that ChatGPT 4o is labelling reviews as "Neutral", unlike BART. We then used ChatGPT 4o to label the set of 30 reviews. Below are the results:

(i) Performance comparison on all 30 reviews

	ChatGPT	**BART**	**VADER**	**TextBlob**
TRUE	26	27	18	21
FALSE	4	3	12	9
Correct_Pred	**86.67%**	**90%**	**60%**	**70%**

Table 7: Performance of ChatGPT 4o on all 30 reviews

(ii) Performance comparison on reviews where either VADER mislabelled, or TextBlob mislabelled, or both

	ChatGPT	**BART**	**VADER**	**TextBlob**
TRUE	13	15	5	7
FALSE	3	1	11	9
Correct_Pred	**81.25%**	**93.75%**	**29.41%**	**41.18%**

Table 8: Performance of ChatGPT 4o on a subset

We can see that ChatGPT is performing better than VADER and TextBlob, and is not performing as well as BART by a slight margin.

NOTE: Our dataset of 30 reviews mostly contains positive reviews, and BART is just labelling reviews as positive or negative. Therefore, the chances of BART getting a label right is higher than other models.
Based on the performance, we created ChatGPT labels as a "Silver-Standard", meaning it is very close and consistent to the label annotations done by our group.

3.2.3 Customer Engagement Forecasting for Businesses

Given below is the classification report for the fine-tuned BERT model:

	precision	recall	f1-score	support
Negative	0.89	0.87	0.88	239
Neutral	0.63	0.63	0.63	298
Positive	0.94	0.94	0.94	1464
accuracy			0.89	2001
macro avg	0.82	0.82	0.82	2001
weighted avg	0.89	0.89	0.89	2001

Table 9: Classification report of Fine-Tuned BERT

We can see that BERT is doing well with classifying Positive and Negative ChatGPT labels, and is struggling with Neutral labels.
To further test the performance of the model before applying it to 658k rows, we applied it on our 30 labelled reviews. Given below are the results:

(i) Performance comparison on all 30 reviews

	BERT*	ChatGPT	BART	VADER	TextBlob
TRUE	24	26	27	18	21
FALSE	6	4	3	12	9
Correct_Pred	**80%**	**86.67%**	**90%**	**60%**	**70%**

** BERT is trained on ChatGPT labels*

Table 10: Performance of Fine-Tuned BERT on all 30 reviews

(ii) Performance comparison on reviews where either VADER mislabelled, or TextBlob mislabelled, or both

	BERT*	ChatGPT	BART	VADER	TextBlob
TRUE	13	13	15	5	7
FALSE	3	3	1	11	9
Correct_Pred	**81.25%**	**81.25%**	**93.75%**	**29.41%**	**41.18%**

** BERT is trained on ChatGPT labels*

Table 11: Performance of Fine-Tuned BERT on a subset

From the performance result, the fine-tuned BERT's performance is very close to ChatGPT, and this indicates that the model is doing well. Therefore, we can apply this to all 658k rows.

Now coming to our XGBoost model for multi-classification of business engagement performance, we used the validation set to see how the accuracy and F1-score are increasing as we increase the training set. Given below are the results:

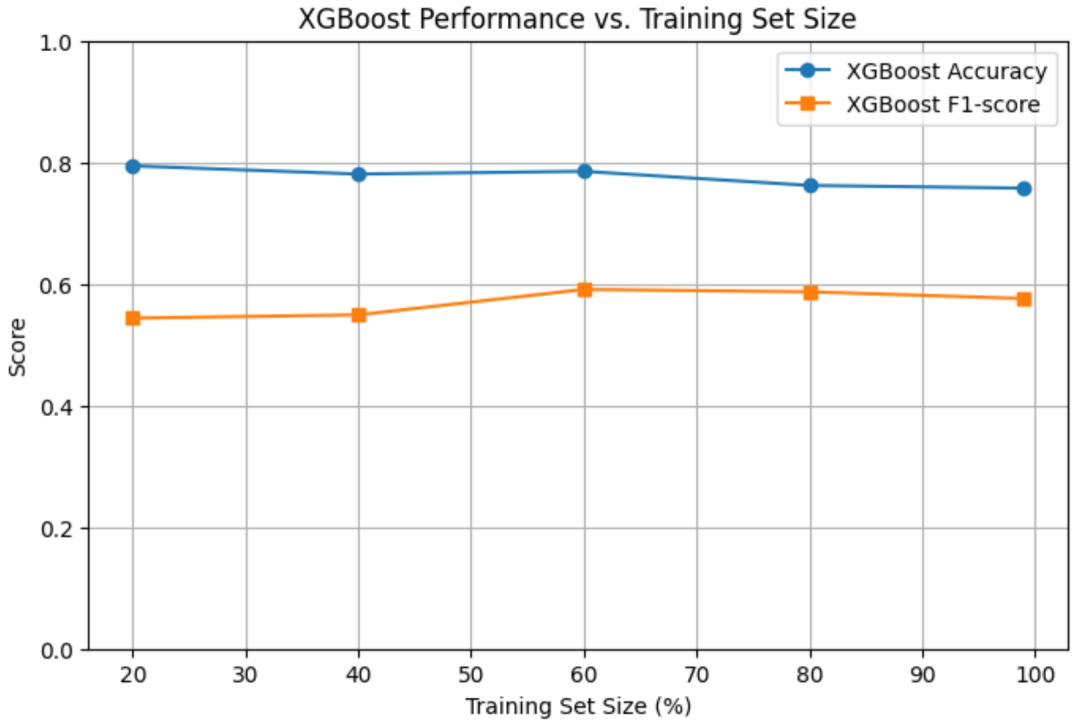


Figure 17: Accurarcy and F1-score.

We then applied the trained XGBoost model on our test set, and we achieved an accuracy of **76%** and an F1-score of **0.58**.
Given below is the classification report:

	precision	**recall**	**f1-score**	**support**
0	0.18	0.32	0.23	85
1	0.90	0.80	0.85	910
2	0.59	0.73	0.65	127
accuracy			0.76	1122
macro avg	0.56	0.62	0.58	1122
weighted avg	0.81	0.76	0.78	1122

Table 12: Classification report with precision, recall, F1-score, and support.

Given below is the confusion matrix. This is to check the true and predicted labels for each class.

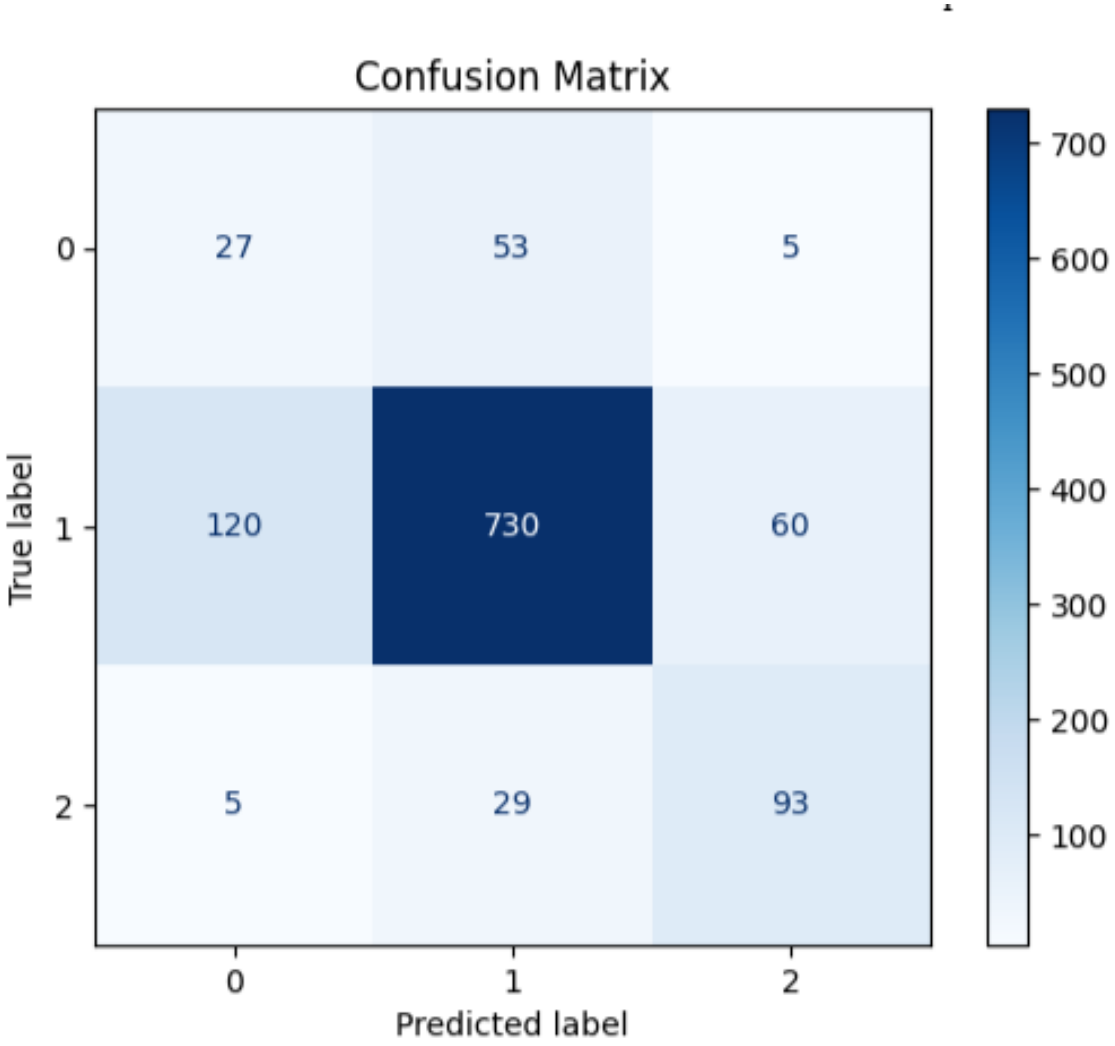


Figure 18: Confusion Matrix of Predicted vs. True XGBoost Labels

Here are the key takeaways from the confusion matrix:

- **Class 1 (Stable)** dominates the test set and is most accurately predicted.
- **Class 0 (Decreased)** is underpredicted — frequently confused with "Stable".
- **Class 2 (Increased)** has moderate precision/recall and is often confused with "Stable".

3.3 Results: K-Means Custering Based on Sentiment, Rating and Volume

These are the 4 clusters we got based on Average rating (1-5), Sentiment label (0-Negative,1-Neutral, 2-Positive) and Number of Reviews per business.

cluster	avg_rating	num_of_reviews	label
0	3.26	131.66	1.36
1	4.43	181.28	2.00
2	4.23	201.05	0.63
3	4.36	3003.97	1.65

Table 13: Cluster-wise summary of average rating, number of reviews, and label value

Cluster Breakdown:

- **Cluster 0:** Businesses with comparatively low ratings, with neutral overall sentiment, and low review volume $->$ likely underperforming or less established businesses.
- **Cluster 1:** Positively rated businesses with positive sentiment and medium review volume $->$ average service and popularity.
- **Cluster 2:** High ratings and positive sentiment but lower review volume $->$ possibly newer or small businesses.
- **Cluster 3:** High-performing businesses with high ratings, positive sentiment, and high review volume $->$ established businesses with probable chain locations nationwide.

To compare how average rating and review volume appears to affect each cluster, we did a simple clustering visualization with Average rating on the Y axis and Number of Reviews on the X-axis color coded by clusters which revealed a clear group separation while also based on the correlation between the three features.

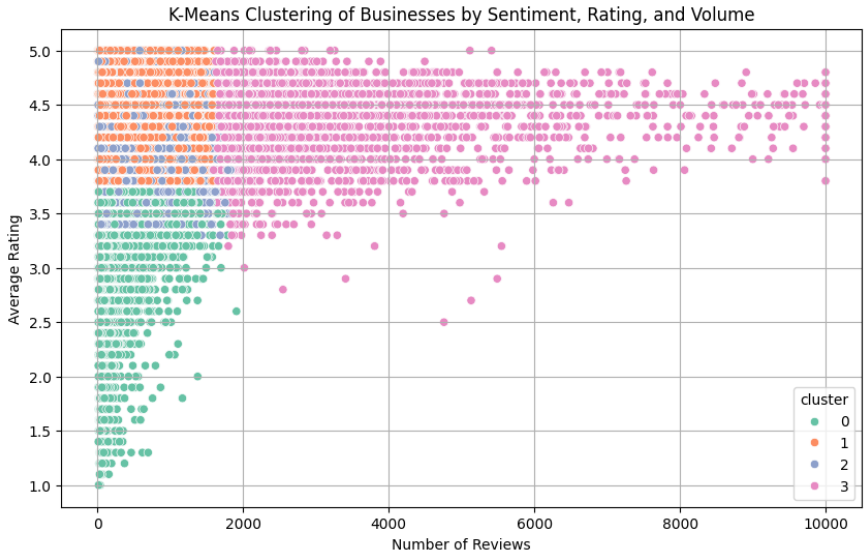


Figure 19: K-Means clustering of businesses based on sentiment label, average rating, and number of reviews.

A 3-D interactive plot (via Plotly) shows the distribution of each cluster based on the three features (Average rating, Sentiment label and Number of Reviews).

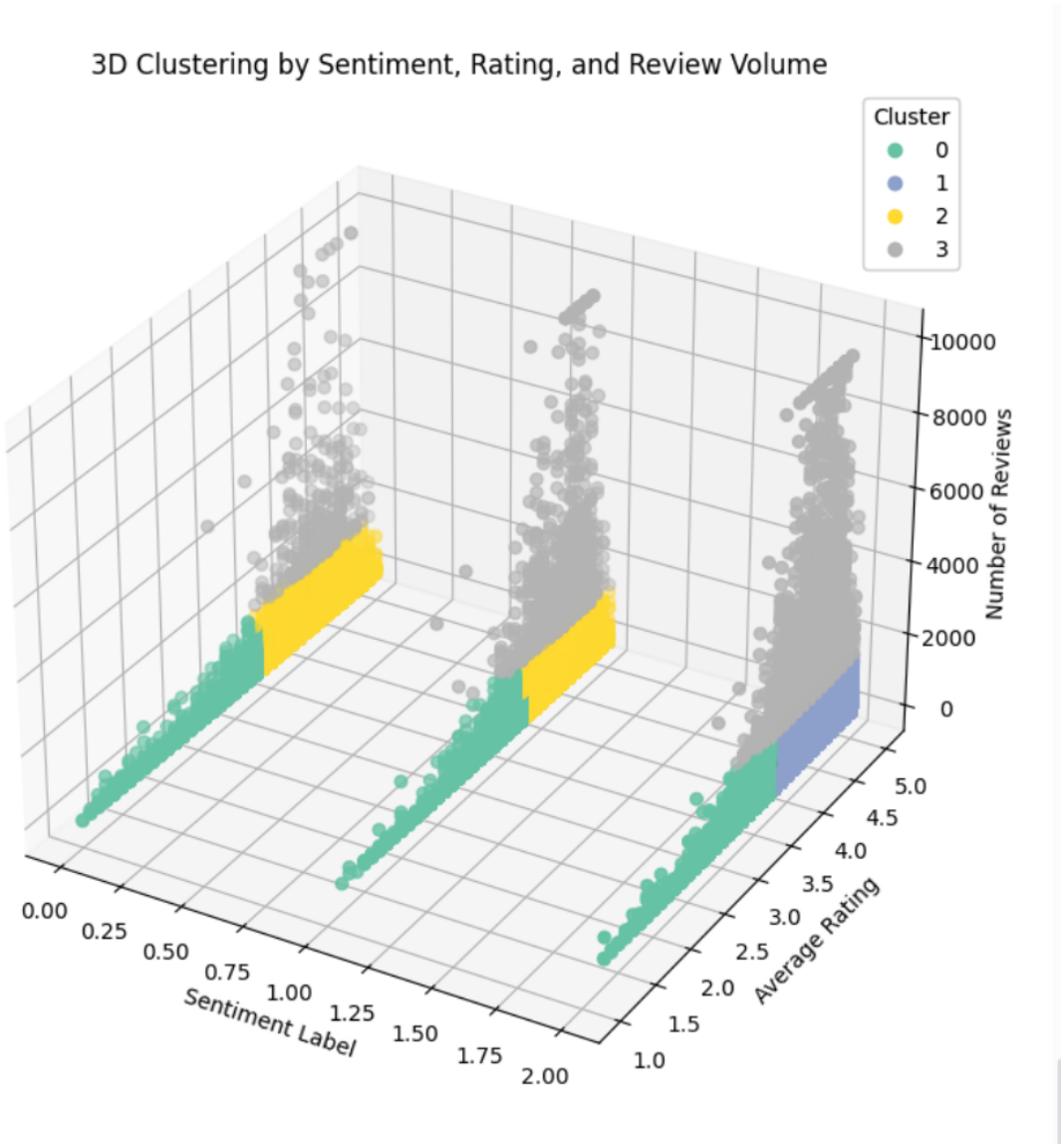


Figure 20: K-Means clustering of businesses based on sentiment label, average rating, and number of reviews.

This plot confirms our interpretations from above while adding interesting insights about the distribution of businesses in all three clusters.

3.4 Results: Business Recommendation Tool - KNN based

To demonstrate the usefulness of the tool, we chose a random business "Qrious Palate" as an example to evaluate the results.

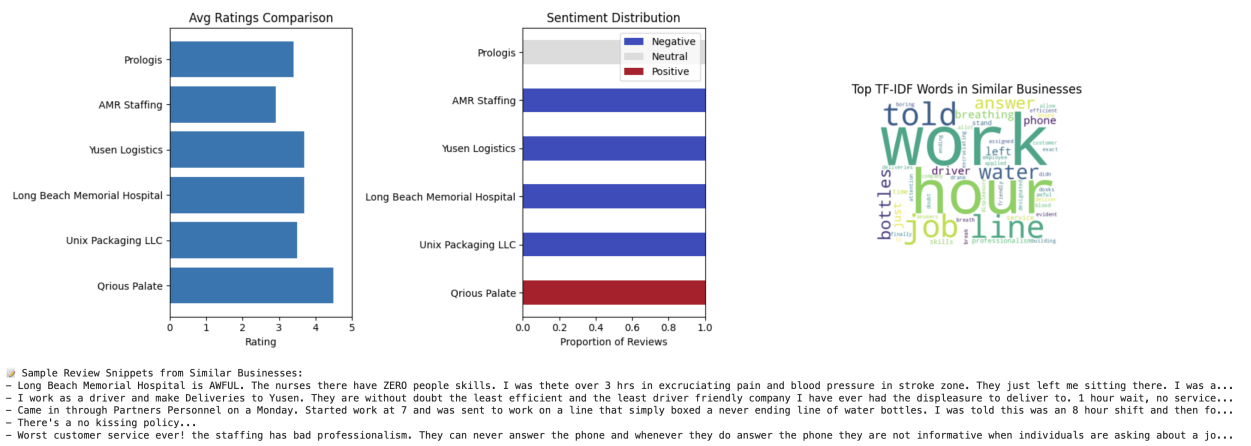


Figure 21: Analysis of business performance and customer sentiment in similar businesses.

The model provided 5 similar businesses with a random snippet of one review from each business. Through this visualization we observed that:

1. Qrious Palate significantly outperformed others in both **average rating** and **positive sentiment share.**

2. Competitors like AMR Staffing and Yusen Logistics had predominantly **negative reviews**, indicating potential service or experience issues.

3. The **word cloud**, constructed using TF-IDF from similar businesses, revealed frequently mentioned and possibly impactful terms such as "job", "work", "hours", and "answer". These reflect operational and customer service topics.

4. In the sample review snippet, most reviews provide a **Negative** or **Neutral** sentiment.

4 Discussion

4.1 Discussion: Sentiment Analysis

The main goal of exploring various sentiment analysis tools for sentiment classification of reviews was to identify a reliable and scalable method. We began with rule-based sentiment tools such as VADER and TextBlob due to their ease of use and lack of training requirements. While both performed reasonably well, our group's annotation experiment on a stratified sample of 30 reviews revealed noticeable limitations. Due to these limitations, we decided to explore zero-shot classification using bart-large-mnli by Facebook. BART performed exceptionally well on our 30-review gold standard, outperforming both rule-based models. But soon we realized an issue, BART was not labelling any review as "Neutral", which became evident when tested on a larger 10k sample. This was the main reason we decided not to scale it over more rows. Ultimately, we opted to use the ChatGPT 4o API for large-scale labeling due to its contextual understanding and consistency. By labeling 10,000 reviews

using stratified sampling, we created a silver-standard dataset that was both diverse and consistent. ChatGPT performed very well on our 30-reviews gold-standard, which reinforced our confidence in its use. Overall, due to the lack of labels to test whether these models were performing well or not for sentiment classification, we explored a variety of tools, which included rule-based as well as advanced transformer based models.

4.2 Discussion : Customer Engagement Forecasting for Businesses

For this part, we focused on forecasting business engagement based on past trends in customer reviews and operations data. The biggest problem we faced was to label about 658,000 reviews for 5608 businesses. When we used ChatGPT API to label our 10k stratified reviews, it was around $2.58 for the total cost. Considering a similar range of token in 658k reviews, we calculated that the total cost to label those reviews will be around $170. Apart from that, when using API, it took almost 40 minutes to label the reviews using API calls and labelling 5 reviews in batches. Therefore, based on the resources we had, it would've taken about 40 hours to label all those 658,000 reviews. This prompted us to look for a more efficient way to label them. So we decided to go with fine-tuning a transformer based model, and we chose DistilBERT because it is lightweight, cheaper, and faster. Evaluation on our gold-standard dataset of 30 reviews confirmed that the fine-tuned model performed comparably to ChatGPT and significantly outperformed rule-based approaches like VADER and TextBlob.
For the XGBoost model, we employed a variety of strategies to tackle class imbalance, but it might not have been enough. The model demonstrated strong performance on the majority class (stable), but struggled more with identifying businesses whose engagement decreased, which can be a likely consequence of class imbalance and fewer negative examples. Ultimately, this modeling approach proved effective at capturing overall engagement patterns, particularly for businesses with consistent review history. While there is room for improvement in minority class performance, the results validate the utility of review-driven features and sentiment signals in forecasting business health over time.

4.3 Discussion: K-Means Clustering Based on Sentiment, Rating and Volume

To better understand the geographical distribution of each cluster, we plotted the clusters onto a heatmap of California using business coordinates. This demonstrated that all clusters follow the same outline around major cities with similar hotspots.

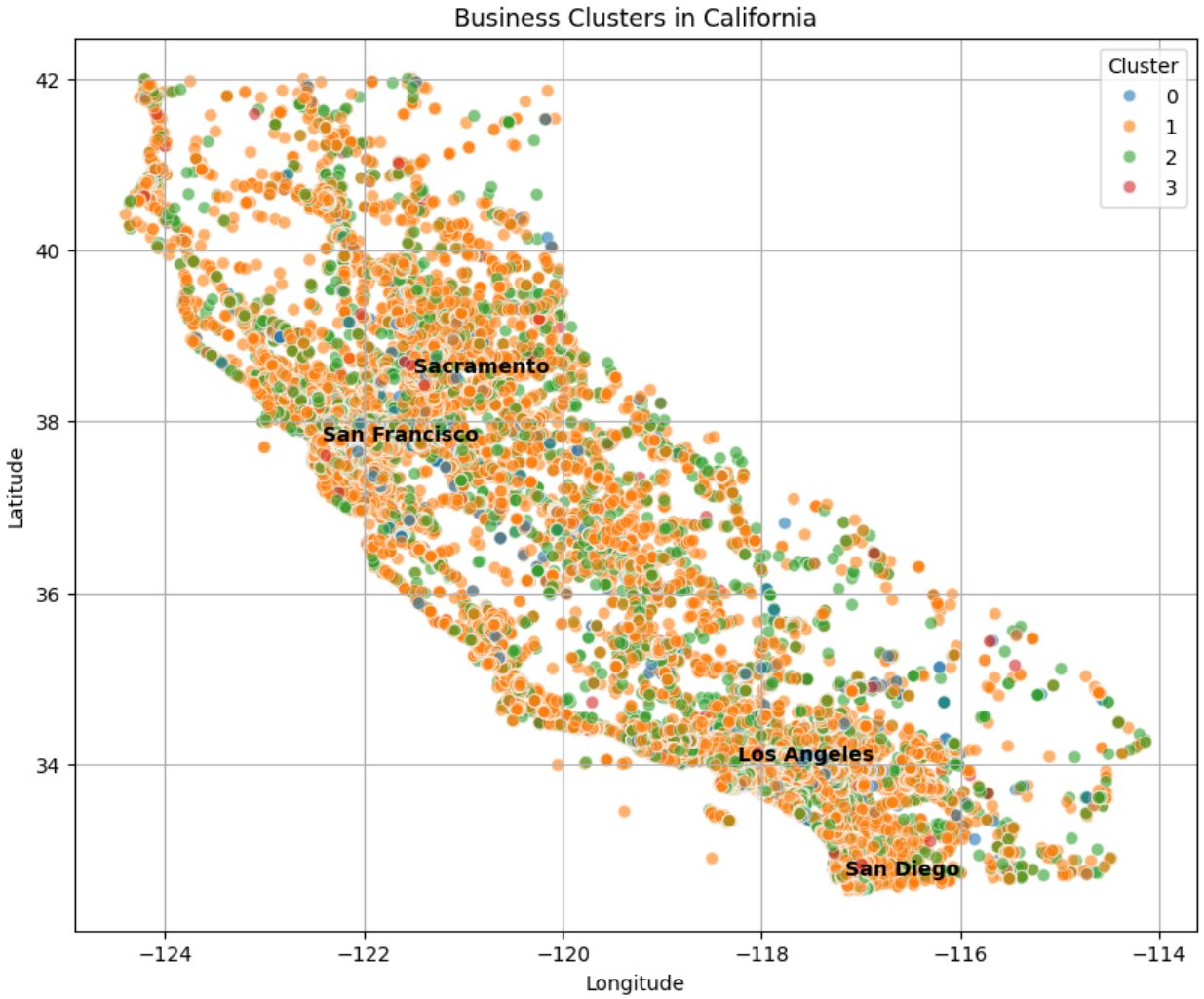
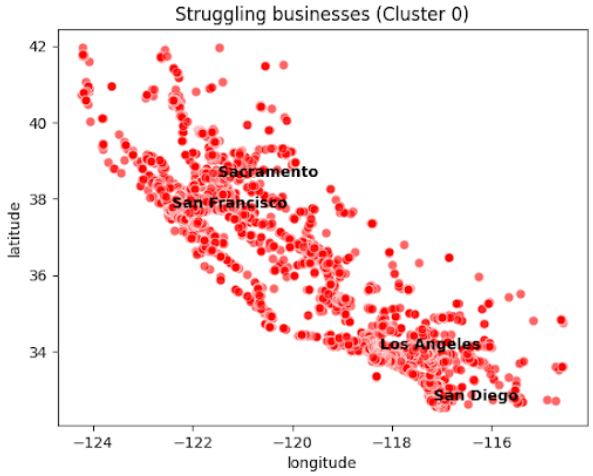
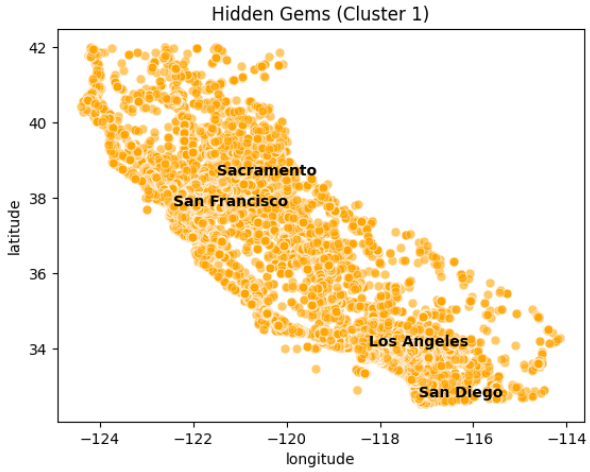


Figure 22: Business clusters across California based on geographic coordinates.

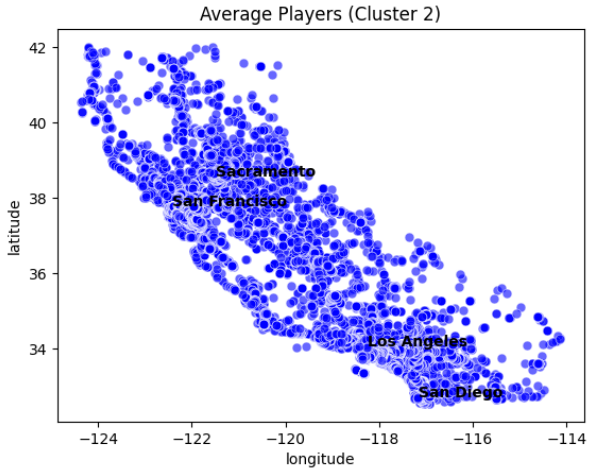
Furthermore to compare the influence of location on each cluster we plotted them separately.



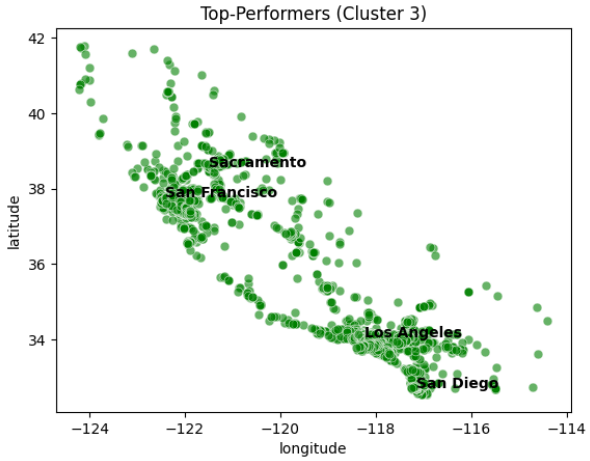
(a) Struggling businesses (Cluster 0)



(b) Hidden Gems (Cluster 1)



(c) Average Players (Cluster 2)



(d) Top-Performers (Cluster 3)

Figure 23: Geographic distribution of businesses in California by cluster type. Each subplot represents a different K-Means cluster.

As we can see most businesses are following the same patterns around cities mostly varying in density but in Cluster 1, there is a broader coverage overall which is interesting indicating booming small businesses in less competitive areas. Clusters with average businesses are spread evenly but Cluster 3 with the top performing business with high popularity are mostly located around the prime urban zones of the state.
To dive even deeper we wanted to evaluate top businesses in our high performing Cluster (Cluster 3) and highlight the areas they are located around using Python's contextily library. These could serve as potential investment worthy areas for business expansion.

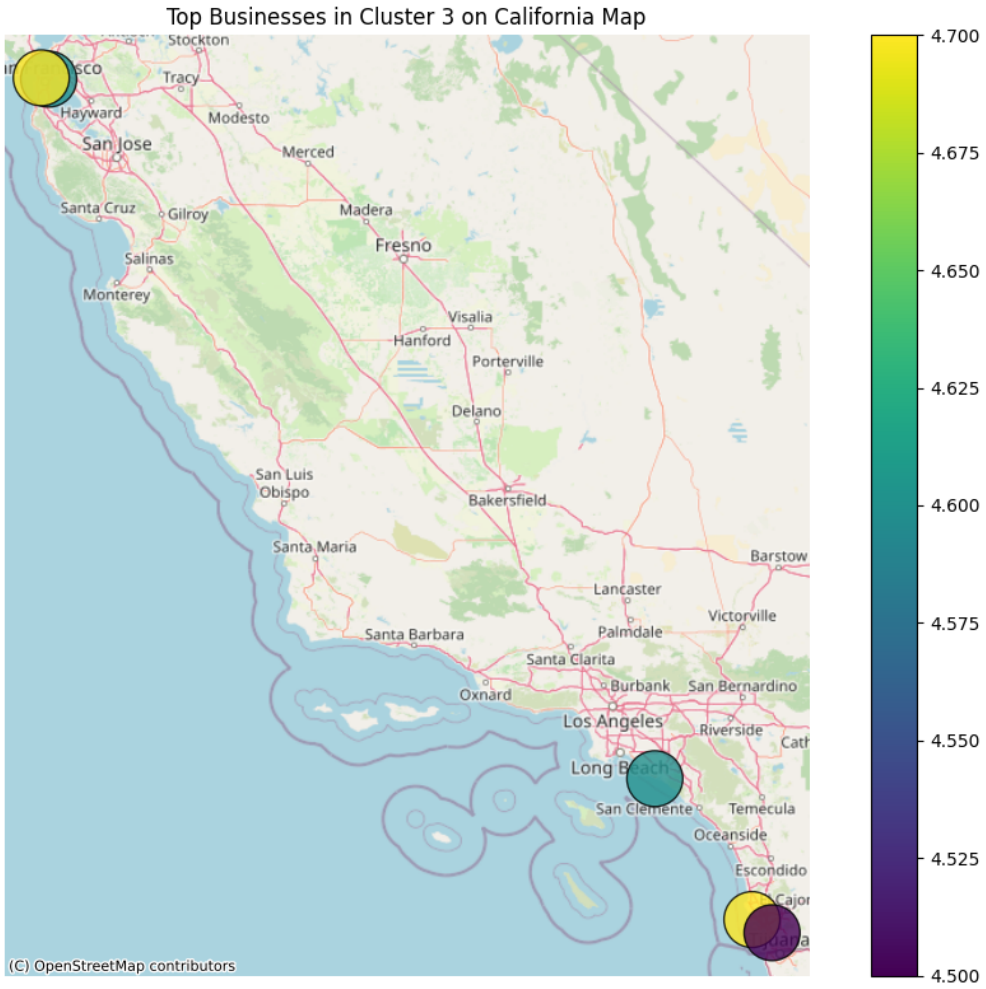


Figure 24: Location and ratings of top businesses in Cluster 3 across California.

4.4 Discussion: Business Recommendation Tool - KNN based

We also made an effort to add tips or improvement suggestions for business owners in our tool. Through spaCy's part-of-speech tagging, we filtered the most actionable keywords from similar businesses' reviews that were missing in the target business's reviews to potential spot themes or expectations that their business might be missing e.g. speed, cleanliness, friendliness etc. This did not perform as expected.

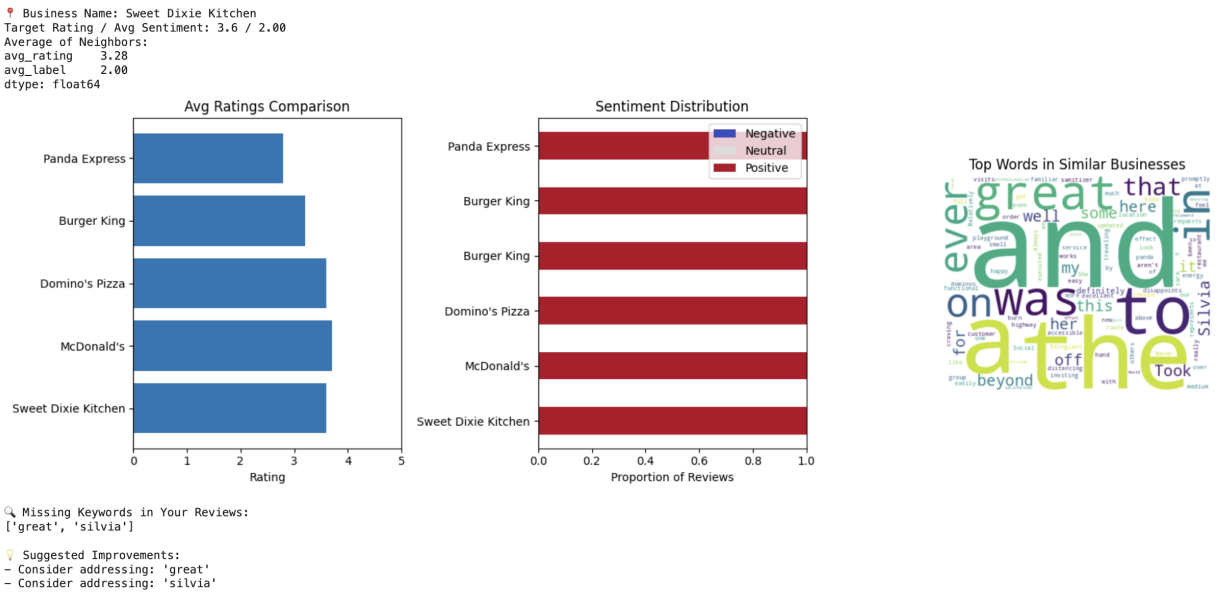


Figure 25: Output

As we can see above the missing words and suggestions are not meaningful while also the word cloud did not necessarily pick any useful words. Due to this challenge, we resorted to using TF-IDF's top words for word cloud and replacing these recommendations with the actual reviews from customers.

5 Conclusion

This project aims to use online customer reviews as a resource to help predict and improve business performance using features like time, location, business type, etc. In this quest to forecast business performance using customer review sentiment and rating, we incorporated multiple NLP techniques ranging from rule based models like VADER and Textblob to Large Language Models like Chat GPT to accurately grasp the sentiment expressed in each review as good as Human Annotation. We also aimed to develop various supervised and unsupervised models like XGBoost for a multi-classification analysis to predict business engagement, which in turn is a direct indicator of business performance, based on historical trends using year-over-year deltas (2018-2019 and 2019-2020) and K-means Clustering to identify and micro analyze state to city level trend among different types businesses using Sentiment, Ratings and Volume as leading factors. In an effort to help businesses be aware of their market and help customers find similar businesses, we aimed to build a recommendation system using K-nearest neighbors where this tool provides insightful graphs comparing the average rating, proportion of Positive, Negative and Neutral Sentiment, Word Cloud with significant action words as well as a snippet of reviews from 5 similar businesses within a 25 km radius. This unique combination of analyzing review sentiment using NLP and machine learning techniques to uncover regional, categorical, and seasonal trends can enable a modern, dynamic system for monitoring business health and forecasting trends, benefiting both businesses and customers by helping them connect with a diverse pool of businesses. We

also acknowledge the under representation of neutral and negative sentiments in customer reviews, likely due to response bias. Additionally, we faced challenges in generating useful recommendations from keywords and recognized the need for a broader baseline to improve review sentiment analysis. Overall, this project presents a scalable, data-driven approach for understanding and enhancing business performance using online reviews, a growing, dynamic, and underutilized resource in the digital age.

References

Alemán Viteri, S. B. (2021). Análisis de sentimientos para twitter con vader y textblob. *Revista Odigos*, 2(3):9–25.

Archak, N., Ghose, A., and Ipeirotis, P. G. (2007). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(07-36).

Bi, J.-W., Liu, Y., Fan, Z.-P., and Zhang, J. (2019). Wisdom of crowds: Conducting importance-performance analysis (ipa) through online reviews. *Tourism Management*, 70:460–478.

Chen, W. and Tabari, S. (2017). A study of negative customer online reviews and managerial responses on social media—case study of the marriott hotel group in beijing. *Journal of Marketing and Consumer Research*, 41:53–64.

Dahiya, A., Gautam, N., and Gautam, P. K. (2021). Data mining methods and techniques for online customer review analysis: A literature review. *Journal of System and Management Sciences*, 11(3):1–26.

Gao, S., Tang, O., Wang, H., and Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management*, 71:19–32.

Giancaterino, C. G. (2023). Zero-shot text classification experience with huggingface. Towards AI, LLM Academy. Last updated on December 31, 2023. Accessed: 2024-04-09.

Guerreiro, J. and Rita, P. (2020). How to predict explicit recommendations in online reviews using text mining and sentiment analysis. *Journal of Hospitality and Tourism Management*, 43:269–272.

Jones, M. (2018). It's all about trust: A brief history of online reviews. Published on WebPunch.

Kumar, N., Kaur, K., Saini, R., Singla, S., and Shilpa (2024). Evaluation of sentiment using deep learning and machine learning using word integration techniques. In *Proceedings of the 2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP)*, pages 278–282, Bali, Indonesia. IEEE.

Kyriakidis, A. and Tsafarakis, S. (2024). Extracting knowledge from customer reviews: An integrated framework for digital platform analytics. *International Transactions in Operational Research*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Li, J., Shang, J., and McAuley, J. (2022). Uctopic: Unsupervised contrastive learning for phrase representations and topic mining.

M., J.-A. (2025). The pros and cons of online customer reviews. Published by IRP Commerce.

Meyer, C. and Schwager, A. (2007). Understanding customer experience. *Harvard Business Review*, 85(2):116–126, 157.

Nowak, A. (2022). The history of online reviews. Published on Expert Reputation.

Pinto, A. S., Pato, M., and Datia, N. (2024). Enhancing drug reviews insights through exploratory data analysis and sentiment analysis. In *Proceedings of the 2024 28th International Conference on Information Visualisation (IV)*, Lisbon, Portugal. IEEE.

Podolsky, M. (2024). Online review trends affecting today's consumers. Published on Forbes Business Council.

Reddy, P. C., Indrani, P., Janaki, P., Gayathri, P., Chandrahasini, P., Apparao, G., and Rajeshwari (2024). Product review sentiment analysis. *International Journal for Multidisciplinary Research (IJFMR)*, 6(3).

Shahare, S. (2022). The history of online reviews and how they have evolved. Published on LinkedIn.

Shen, W. (2008). Essays on online reviews: The relationships between reviewers, reviews, and product sales, and the temporal patterns of online reviews.

Sprague, D. D. (2025). The history of online reviews and how they have evolved. Originally published: December 20, 2019.

Tao, S. and Kim, H.-S. (2022). Online customer reviews: Insights from the coffee shops industry and the moderating effect of business types. *Tourism Review*, 78(3):789–807.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Yan, A., He, Z., Li, J., Zhang, T., and McAuley, J. (2023). Personalized showcases: Generating multi-modal explanations for recommendations.

Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Appendix A:
Project Code: https://github.com/Anshul-Kum/DAT490