CUSTOMER SEGMENTATION USING SENTIMENT ANALYSIS

Table of Contents

Customer Segmentation Using Sentiment Analysis on Yelp Restaurant Reviews

Abstract:

With the growing importance of online reviews in the restaurant industry, understanding customer sentiment is crucial for business improvement. This project leverages natural language processing (NLP) to analyze Yelp reviews of restaurants and group customers based on their sentiment using clustering techniques. The result is a classification of customers into segments: Positive, Neutral, and Negative, which helps businesses identify satisfaction levels and take targeted actions.

Keywords:

Sentiment Analysis, Customer Segmentation, Yelp Dataset, Natural Language Processing (NLP), TextBlob, KMeans Clustering, Restaurant Reviews, Customer Feedback Analysis, Online Reviews, Business Intelligence, Unsupervised Learning, Opinion Mining, Customer Satisfaction, Data-Driven Marketing, Consumer Behavior, Polarity Score, Text Mining, Review Analytics, NLP in Business, Customer Insights.

Chapter 1: Introduction

Introduction:

In the digital age, customers often share their dining experiences on review platforms like Yelp. These reviews provide valuable insights into customer satisfaction, food quality, service, and ambiance. However, due to the sheer volume of reviews, it becomes difficult for restaurant owners or managers to manually read and understand all of them.

This project provides a solution by automating the analysis of Yelp reviews. We use sentiment analysis to quantify the emotional tone of each review and apply clustering to group customers into sentiment-based segments. This can help businesses:

- Recognize loyal and happy customers,
- Identify dissatisfied customers early,
- Tailor marketing and service strategies accordingly.

Problem Statement:
Restaurants receive a massive number of customer reviews daily, and there is no efficient manual method to extract useful information from them in a structured form.

The problems we aim to solve:

- How can we automatically understand customer opinions at scale?
- Can we group customers based on how they feel?
- How do we visualize and interpret these groups in a useful way?

Our objective is to create a pipeline that:

1. Extracts restaurant reviews.
2. Analyzes the sentiment of each review.
3. Segments the customers into categories based on the tone of their feedback.

Chapter 2: Literature Review

Sentimental Analysis on Twitter Data:

The growing importance of online platforms has fueled extensive research in the domain of sentiment analysis, particularly using user-generated content from social media sites like Twitter and Yelp. These platforms provide massive volumes of informal, opinion-rich text data, which has become a valuable resource for understanding public sentiment and consumer behavior.

In the paper *"Sentiment Analysis on Twitter Data"* (Sahayak et al., 2015), the authors propose a machine learning-based sentiment classification framework to analyze tweets. Their work emphasizes the importance of extracting opinions from microblogs, classifying them as positive, negative, or neutral using algorithms such as Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM). The study also highlights the use of noisy labels such as emoticons and acronyms for training sentiment classifiers—a technique particularly effective in short-form text like tweets.

Although Twitter and Yelp differ in context—tweets being real-time and spontaneous while Yelp reviews are often more reflective and detailed—the underlying principle of sentiment extraction remains common. Both types of content benefit from text preprocessing steps such as tokenization, stop-word removal, and handling of negations, which are essential to improving the accuracy of NLP models.

The Twitter study also introduces models like unigrams, tree kernels, and feature-based classifiers. Similarly, in our Yelp review analysis, sentiment polarity scores are calculated using TextBlob (which internally uses a unigram model), and customer segmentation is achieved through KMeans clustering. While the referenced work employs supervised learning with labeled training data, our approach uses unsupervised learning (clustering) to segment users based on sentiment without prior labels, making it more adaptive for new datasets.

Moreover, the literature suggests that sentiment classification is highly domain-specific. Algorithms fine-tuned for Twitter may not perform well on Yelp without adaptation, due to differences in language style, length, and intent. For example, while tweets are limited in characters and often include slang or hashtags, Yelp reviews are longer, more descriptive, and structured. This distinction is critical in determining the feature engineering and preprocessing required for effective sentiment analysis.

In conclusion, existing literature on Twitter sentiment analysis offers valuable insights and methodologies that can be adapted for Yelp review analysis. Techniques such as feature extraction, sentiment scoring, and classification are transferable, although modifications are needed to suit the dataset and domain. Our project builds upon this foundation by applying sentiment analysis to Yelp restaurant reviews and using unsupervised learning to uncover customer segments, thereby providing practical, data-driven insights to businesses in the hospitality industry.

Sentimental Analysis Methodologies and Practices:

Sentiment analysis has gained significant attention over the past decade due to the exponential growth of user-generated content on social platforms like Twitter, Facebook, and Yelp. Researchers have explored various methodologies to mine this data and understand public opinion, customer satisfaction, and behavioral patterns. According to Sahayak et al. (2015), sentiment classification on platforms like Twitter involves extracting text features and using supervised machine learning models such as Naive Bayes, Maximum Entropy, and Support Vector Machines to classify sentiments into positive, negative, or neutral categories. Their study demonstrated that even short, noisy data like tweets could be effectively mined using basic lexical features combined with preprocessing and model tuning.

Building upon this, Patel et al. in *“A Review on Sentiment Analysis Methodologies, Practices and Applications”* (2020) provide a comprehensive overview of techniques used in sentiment analysis. The paper categorizes sentiment analysis into document-level, sentence-level, and aspect-level analysis and emphasizes the importance of natural language processing (NLP) in each. The authors highlight the growing shift from traditional machine learning models to deep learning approaches such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and hybrid models for greater accuracy and context understanding. However, they also acknowledge that simpler tools like TextBlob and VADER continue to provide competitive results for large-scale text analysis due to their speed and interpretability.

In the context of review platforms like Yelp, sentiment analysis not only helps in understanding customer opinions but also supports downstream applications such as customer segmentation, recommendation systems, and business performance analysis. While most academic works focus on classification, relatively fewer studies apply clustering techniques like KMeans to segment users based on sentiment patterns. This gap provides an opportunity for innovative applications that combine unsupervised learning and opinion mining.

Our project leverages this approach by using TextBlob to assign sentiment polarity scores to Yelp restaurant reviews and then applying KMeans clustering to identify customer segments. Unlike studies that rely solely on labeled datasets, our method is unsupervised and scalable, making it suitable for continuous feedback monitoring. By visualizing sentiment distributions and customer groups, our work extends the practical application of sentiment analysis beyond classification and toward strategic business intelligence.

Overall, existing literature affirms the value of sentiment analysis in understanding public opinion and consumer behavior. It also encourages combining NLP with both supervised and unsupervised learning techniques, as done in this project, to gain deeper, actionable insights from customer reviews.

Chapter 3: Methodology

Data Understanding

We used the Yelp Open Academic Dataset, which contains:

- Business details: including business IDs, names, and categories.
- Reviews: customer-written feedback with star ratings.

From the dataset, we focused on:

- Businesses tagged under the category "Restaurants"
- Reviews written for these businesses

Business Understanding

1. Data Filtering

- We extracted all businesses from the business.json file.
- From these, we kept only those categorized as "Restaurants".
- Then we loaded review data from the review.json file and filtered reviews that matched the restaurant business IDs.

2. Sentiment Analysis

- We used the TextBlob library to compute sentiment polarity for each review.
- The polarity score is a float ranging from -1 (very negative) to +1 (very positive).
- Example:
  - "Amazing food!" → Polarity: 0.8 (Positive)
  - "Just okay." → Polarity: 0.2 (Neutral)
  - "Horrible service." → Polarity: -0.6 (Negative)

3. *Customer Segmentation (Clustering*)

- We used the KMeans clustering algorithm to group reviews into three clusters.
- The clustering was performed based only on sentiment polarity.
- After clustering, we manually assigned labels to each cluster based on their average polarity: Negative, Neutral, Positive.

4. *Visualization*

- A histogram was created to show the distribution of sentiment scores across all reviews.
- The bars were color-coded by segment (cluster label) to easily observe how sentiment is distributed.

Chapter 5: Results and Findings

In this study, a total of 20,000 Yelp restaurant reviews were processed to perform sentiment-based customer segmentation. Each review was analyzed using sentiment polarity scores generated by the TextBlob library, which ranged from -1 (very negative) to +1 (very positive). These scores were then used to cluster the reviews into three distinct sentiment groups, Negative, Neutral, and Positive using the KMeans clustering algorithm.

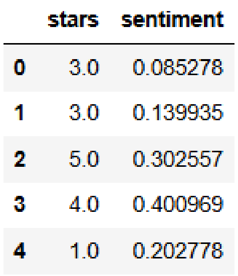
Upon analyzing the clustering results, it was observed that the majority of reviews fell within the neutral sentiment range, with polarity scores approximately between 0.1 and 0.3. This indicates that most customers had a moderate experience not highly dissatisfied but not overly enthusiastic either. These reviews often included mixed or balanced opinions, such as average food quality or inconsistent service, reflecting a middle-ground sentiment.

A smaller segment of reviews exhibited high sentiment polarity, typically above 0.5, and were thus grouped into the positive cluster. These reviews often expressed strong satisfaction, highlighting excellent food, great service, and pleasant ambiance. Customers in this segment are likely to become repeat visitors and brand advocates.

Conversely, a narrow segment of reviews had sentiment scores below 0.1, identifying them as part of the negative cluster. These reviews included complaints about poor customer service, low food quality, or bad experiences. Although this group was the smallest, it holds strategic importance, as identifying and addressing their concerns can significantly improve customer retention and brand reputation.

Overall, the sentiment-based clustering successfully highlighted how customers perceive their restaurant experiences and provided a structured framework for identifying different levels of satisfaction among Yelp users

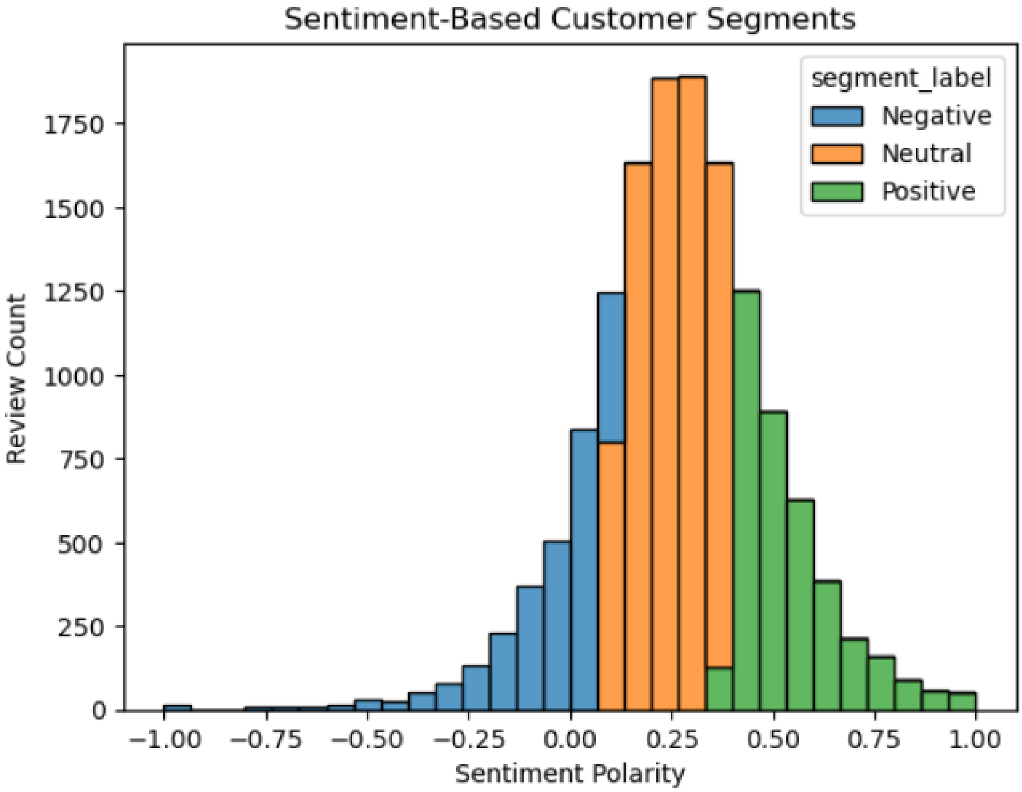
Output



	stars	sentiment
0	3.0	0.085278
1	3.0	0.139935
2	5.0	0.302557
3	4.0	0.400969
4	1.0	0.202778

Histogram Insight

- The majority of sentiment scores clustered around neutral values (0.1 to 0.3).
- A smaller group of reviews had high sentiment (above 0.5) and were labeled positive.
- A narrow segment of reviews had sentiment scores below 0.1 and were considered negative.



Observations

- Many reviews labeled with 3 stars had mixed sentiments, landing in the neutral category.
- Reviews with high star ratings (4-5) mostly fell in the positive sentiment group.
- Reviews with 1-2 stars usually had negative or neutral sentiment.

Conclusion

This project successfully demonstrates how natural language processing and clustering can be used to analyze and categorize customer feedback at scale. The sentiment-based segmentation approach offers several benefits to restaurants:

- Improve service by identifying negative feedback patterns.
- Retain customers by following up with neutral or dissatisfied customers.
- Encourage loyal customers to become brand advocates.

This approach not only saves time but also empowers data-driven decision-making in the food service industry. With minor enhancements, this model can also:

- Track sentiment trends over time,
- Analyze by location or business type,
- Integrate with recommendation systems.

References:

[1] Sahayak, V., Shete, V., Pathan, A., & Department of Information Technology, Savitribai Phule Pune University, Pune, India. (2015). Sentiment analysis on Twitter data. In *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* (Vol. 2, Issue 1). https://d1wqtxts1xzle7.cloudfront.net/36584429/28.JACS10092-libre.pdf?1423574431=&response-content-disposition=inline%3B+filename%3DSentiment_Analysis_on_Twitter_Data.pdf&Expires=1724869234&Signature=awf68RwOym7NsLD-s3X2e3jhlgLmzw49jYX3VNhDjqQIaysaN4JaJjucVwj6wuhl~YnPIyYVQF2eeBXVI2vRbXqnORI4T7LBvJZAVSB6J3Xnbi6aVX5uvN7PfP04vxmasIjZG-Zuxpg3UFliNn-UJmfwv5apTEbLA~t4FGn8DD-86IDxNQwZdwU7w0IDAVBE1OR-m8z70x1c7qlITjVbhKlJUoRklMcZCR3syhjsiL1tVtWKN6-CLD8FtxGfrPNP3nB8TZA6jrbB4sD~6plqVilNuzKrgyC066cscD0eqQxaYhLBikivzzNPCt20EAwCbs0uOY5Mgqx~esxpCEDmgA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[2] Mehta, P., & Pandya, S. (2020). A review on sentiment analysis Methodologies, Practices and applications. In INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH* (Vol. 9, Issue 02, pp. 601–602) [Journal-article]. https://www.researchgate.net/profile/Pooja-Mehta-26/publication/344487215_A_Review_On_Sentiment_Analysis_Methodologies_Practices_And_Applications/links/5f7bfb2992851c14bcb16528/A-Review-On-Sentiment-Analysis-Methodologies-Practices-And-Applications.pdf