Group 1: Sentiment Analysis and Business Rating Predictions from Google Local Reviews

Maddy Chapnik-20270067, Joanna Bian-20149892,
Alex Marinkovich-20347130, Lucas Adams-20372198

1 Motivation and Problem Statement

Online reviews are a valuable tool that can be used to determine consumer sentiment and to gain business insights. Using the Google Local Reviews dataset, this project aims to determine common themes in online reviews across various business types, analyze variations in consumer sentiment across different geographical and business areas, and develop a model to predict business ratings based on textual review content. The input to the project is the Google Local Data (2021) dataset, which contains useful data such as business name, location, review text, user rating, and more. The expected outcomes of the models include insights into themes in customer reviews, sentiment distribution trends, and a predictive model for business ratings based on review texts. To measure the success of the project, accuracy and error metrics will be used to determine the validity of the models, while statistical significance tests, visualization methods and qualitative analysis will be used to help interpret findings.

Online reviews play an essential part in shaping consumer perceptions, and hence can be a significant contributor to business success or failure. Analyzing online reviews allows businesses to gain an understanding of customer sentiment, identify strengths and areas for improvement, and create actionable plans to implement positive change. However, without access to proper data analysis and predictive techniques, businesses may struggle to extract valuable insights from large datasets. As such, the models created within this project can become valuable tools to help business owners grow their operations effectively. The Google Local Reviews dataset encompasses a wide variety of business types and geographic locations, which should help ensure that the models can be generalized to provide insights for a wide variety of companies.

This project will make novel contributions to the realm of sentiment analysis within online reviews. By analyzing consumer sentiment variations across a variety of both business categories and geographical regions, this project will provide a broad range of conclusions that may not have been previously explored together. Furthermore, most sentiment analysis studies are categorical, attempting to assign a sentiment to a body of text (positive, neutral, negative, etc..) [2]. This project, however, will use regression to assign the text a rating between 1 and 5 stars, and will investigate which factors contribute most strongly to these ratings.

2 Research Questions and Methodology

Three research questions will be investigated in this project.

2.1 RQ1: What are the dominant themes in customer reviews, and how do they correlate with star ratings?

Motivation: Understanding the dominant themes in customer reviews is essential for business to identify strengths and weaknesses based on customer reviews. By analyzing how these themes correlate with star ratings, businesses can gain insights of different aspects that lead to positive and negative reviews. This is very valuable information as it can be used for making strategic decisions to modify or improve on services provided.
Proposed Methodology: Data preprocessing will be first used to extract review texts and corresponding star ratings from the dataset. Then we will develop a regression model to understand how different factors influence star ratings. After that, pre-trained BERT-based models will be used to extract deep contextual meaning from the review texts. And finally, we will perform a correlation analysis between extracted themes and review ratings to understand which themes are associated with positive or negative reviews.

2.2 RQ2: How does sentiment vary across different geographical regions and business categories?

Motivation: Researching how sentiment varies across geographical regions and business categories provides valuable insights for businesses, helping them make informed decisions. Using trends in sentiment analysis, businesses can tailor their strategies to enhance the customer experience. For example, a restaurant chain may find that customers in urban areas leave worse reviews due to longer wait times, while in suburban areas, negative reviews may be tied to having less food options. Based on these insights, the chain can introduce faster service models in urban areas, and an expanded menu for suburban areas.
Proposed Methodology: The research for this question involves data preprocessing, involving converting categorical variables (business category, price level) into numerical representations. Next, sentiment analysis scores will be created for each review using an NLP model, selected based on

its performance and effectiveness. Also, geographical and business category trends will be visualized using statistical methods and heatmaps. Lastly, predictive modeling will include sentiment classification using regression, and deep learning to identify the factors that influence sentiment across regions and industries. Additional techniques like clustering and topic modeling may help identify patterns as well.

2.3 RQ3: Can we accurately predict business ratings based on review text?

Motivation: On a platform like Google, local customer reviews provide valuable insights into how a business performs and their customer satisfaction. However, online business ratings are often biased due to personal biases and external circumstances. By using Natural Language Processing (NLP) and machine learning models, our aim is to find out if we can rely on textual reviews alone to predict a businesses rating. If successful in doing so, this could give businesses real-time input on customers' feedback and opinions, allowing them to anticipate rating trends and overall improve their services.

Proposed Methodology:

1. Data Preprocessing and Cleaning:

Remove noise in text reviews like special characters, numbers, and unnecessary symbols. Split text into individual units(Tokenize) and remove stop words like "the" or "and" so text is readable for NLP. Turn all text into lowercase(lemmatization) for text normalization.

2. Feature Engineering and Sentiment Analysis:

Convert text data into numerical representation by using Term frequency to measure how frequently a word is used in a review and how rare that word is compared to all reviews in the dataset. Then use sentiment analysis to associate certain terms with positive, negative or neutral reviews.

3. Predictive Modeling:

Train machine learning models (e.g., Linear Regression, Random Forest, XGBoost) and deep learning models (e.g., LSTMs, BERT-based classifiers) to predict business ratings based on review text. Then Compare model performance using the evaluation of metrics such as the Root Mean Square Error and classification accuracy.

3 Dataset

The dataset to be used in this project is the "Google Local Reviews" dataset. This dataset was published in 2021 by the UCSD McAuley Group [3]. The dataset contains review information from Google Maps as well as metadata on the businesses themselves, up to September 2021 in the United States. The entire dataset consists of 666,324,103 reviews, 113,643,107 users and 4,963,111 businesses. The dataset has been subdivided by states and has been reduced for practical purposes such that each of the users and items listed have a maximum of 10 reviews each.

The review information portion of the dataset contains the following information that we plan to focus on:

- user id: reviewer ID
- time: time of the review (unix time)
- rating: rating given to the business (star rating from 1-5)
- text: text included in the review
- gmap id: business ID

The business metadata contains the following information that we plan to focus on:

- name: name of the business
- address: address of the business
- gmap id: business ID
- category: business category
- avg rating: average rating of the business
- num of reviews: number of reviews

4 Group Member Contributions

Each group member's contributions to this report are as follows:

Report Section	Group Member
Motivation and Problem Statement	Maddy
Research Question 1	Joanna
Research Question 2	Alex
Research Question 3	Lucas
Dataset	Maddy

References

[1] E. Bigne. Are customer star ratings and sentiments aligned? a deep learning study of the customer service experience in tourism destinations - service business. SpringerLink, 2023.

[2] A. A. Z. H. Huang and M. B. Mustafa. Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. In *IEEE Access, vol. 11*, page 90367–90382. IEEE, 2023.

[3] J. L. Tianyang Zhang. Google local review data. University of California San Diego, 2023.