BA820 – Project Proposal

· **Categorizing and Simplifying Reviews on Yelp**

· **Section A1, Team 10**

· **Aryan Sehgal, Barrett Ratzlaff, Manyi Hong, Shreya Lodha**

· **February 3rd, 2025**

Business Relevance

Customers usually would like to know the opinions of people who have already made the purchase before they make purchasing decisions, therefore, companies like Yelp, Tripadvisor have emerged to offer platforms for customers to see reviews on their target services. However, these platforms face significant challenges that hinder user satisfaction and decision-making:

1. **Hard to align reviews with users' preferences.** For example, some users care about the atmosphere of a restaurant, whereas others prioritize good service much more than the atmosphere. Parsing through reviews to find information that is meaningful to the user can be difficult. It has been found that going through review after review can cause more confusion than certainty.[1]
2. **Difficult to distinguish genuine reviews from unhelpful reviews.** Users may suspect a review is untruthful or may be hesitant after reading one that deviates from every other review. What's more, the proliferation of fake reviews can also undermine user trust in review platforms.

Project Motivation

Simplifying the process of understanding what other patrons of a business think, while ensuring the information is genuine and relevant to users' specific interests, serves as the primary motivation for our project. Yelp has already begun to implement A.I. to categorize the reviews of a business,[2] showing that there is value in making the user experience easier on review platforms. **By applying unsupervised machine learning techniques on textual data (reviews). Our goals are:**

1. **Summarize information and assign it to categories relevant to users**
2. **Evaluate the content quality of each piece of information**
3. **Provide users with prioritized recommendations based on their preferences and the reliability of the reviews.**

This approach not only is time-saving for customers but also helps businesses better understand customer feedback beyond average star ratings.

Problem Statement

Through applying clustering and natural language processing on the text data made available by Yelp, we will simplify the complicated process of finding relevant information for users. Firstly, we will need to identify different "themes" of reviews with unsupervised machine learning methods. We will need to use unstructured data methods on the reviews, because

[1] Ann Kronrod and Yakov Bart. "The Paradox of Language Repetition in Product Reviews." SSRN Electronic Journal (2018). https://doi.org/10.2139/ssrn.3194326.

[2] Jess Weatherbed. "Yelp's new AI-powered review filters will show what you want to know." *The Verge*, December 10, 2024. https://www.theverge.com/2024/12/10/24317749/yelp-ai-powered-review-insights-filters-availability.

leaving the reviews as it is would make it difficult to categorize effectively. Unsupervised machine learning methods like k-means clustering will be necessary because in order to effectively cluster the data, we will need to create dimensions based on the content of the text and the sentiment detected.

In addition to segmentation, we will also evaluate the content quality of each review by using the TF-IDF[3] method. By assigning scores, we can prioritize reviews for users, ensuring that the most reliable and relevant reviews in that category are presented first. If we only clustered based on length, votes, and other dimensions that come with the dataset, we would not be able to generate meaningful summaries from those clusters. A successful project would have two things: An informative segmentation of reviews and a way to summarize the information contained in each segmentation.

Dataset

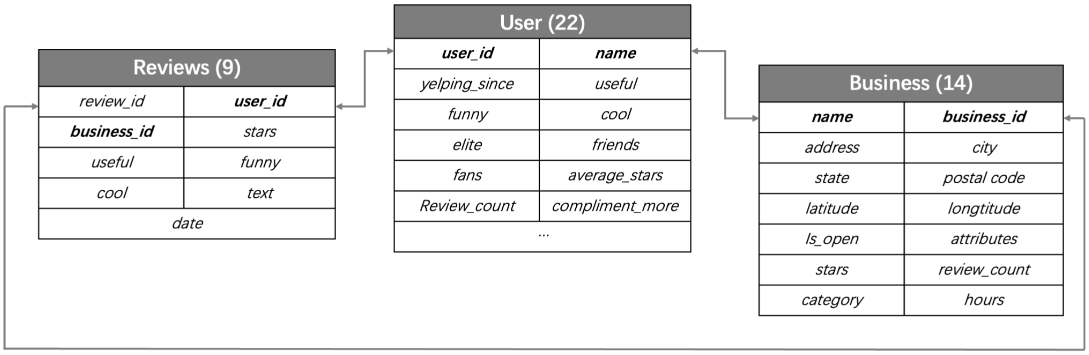
Data source

Our dataset is offered publicly by Yelp, which provides real-world data related to businesses including reviews, photos, check-ins, and attributes like hours, parking availability, and ambiance. https://business.yelp.com/data/resources/open-dataset/

Data description

This dataset contains 1 PDF (data directory and description) and 5 JSON files (8.65 GB), including *business.json*, *checkin.json*, *review.json*, *tip.json*, *and user.json*, with features such as business details, user information, reviews, and ratings. Key features include textual review data, star ratings, and metadata like business categories and location.

Key tables relationship:



Reviews (9)	
review_id	***user_id***
business_id	*stars*
useful	*funny*
cool	*text*
date	

User (22)	
user_id	***name***
yelping_since	*useful*
funny	*cool*
elite	*friends*
fans	*average_stars*
Review_count	*compliment_more*
...	

Business (14)	
name	***business_id***
address	*city*
state	*postal code*
latitude	*longtitude*
ls_open	*attributes*
stars	*review_count*
category	*hours*

The key features we plan to work with are in the *reviews.json*, *user.json*, and *business.json* files. The centerpiece of our project is the "text" feature in the reviews.json, as that is what we plan to categorize and evaluate. We also plan to iterate this process per business, so the "business_id" feature will be important as well. "stars" will be a component of our evaluation of the text_data,

[3] Mamata Das, Selvakumar K., P.J.A. Alphonse. "A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset." (2023). https://arxiv.org/abs/2308.04037

and “user_id” will be used to connect user data to a review so we can further evaluate the content of a review.

Computational Efficiency

To manage our datasets large size, we plan to:

1. Sample a subset of the data: The JSON files together total about 8 gigabytes in storage, sample a subset can save budgets.
2. Remove irrelevant features: Some features, like “postal_code”, are not expected to be informative.
3. Create new features: We expect to engineer a few more features based on the review itself, like length or word count, from the review feature to make our clustering more meaningful.
4. Utilize BU’s Computing Cluster or Cloud Computing: Should sampling not give us sufficient information to work with, we can upload a larger portion or all of the file to cloud storage and perform our project with some form of cloud computing.

Proposed Methodologies

During the process, we would use Natural Language Processing methods like Topic Modeling and Sentiment Analysis, Clustering method, and Summarization method.

For the first step in the project, we will apply Topic Modeling to identify themes in the selected reviews. For example, for a restaurant, we would hope to identify common topics of reviews, like service or pricing. While the previous example might be the expectation, we will avoid imposing our will on the dataset. If unexpected, valid themes appear, we will not continue to cluster until we get the exact themes that we expect. After grouping reviews by subject, we plan to apply sentiment analysis to get a sense of general opinion on that aspect of any one business. TF-IDF & sentiment analysis are important as means to determine if popular opinion on that topic is mostly positive or negative. This is a key piece to generating a meaningful summary for the user.

After the reviews for any one business are segmented sufficiently and overall sentiment is determined within each group, we can focus on generating a summary or suggesting a “most relevant” review. To achieve this, we hope to use a simplified version of retrieval augmented generation (RAG), which is explained in the article in this page’s footnote.[4] Depending on our time constraints, what we employ could range from leveraging LLMs to generate a summary based on the review data within a certain category for a business, or simply pulling the review that is most similar or relevant to that category. After this project, the process could theoretically be integrated into a review platform’s website and be personalized based on profile settings. For example, if a user selected “Atmosphere” as an aspect of interest, one of the first things they would see when visiting a business’s page is either a summary of atmosphere-related reviews or simply an atmosphere-related review that represents the majority opinion recorded.

[4] Learn by Building AI, "A Beginner's Guide to Building a Retrieval Augmented Generation (RAG) Application from Scratch," *Learn by Building AI*, accessed February 3, 2025, https://learnbybuilding.ai/tutorials/rag-from-scratch.

Appendix

01 References

1. Afrimi, Daniel. *"Text Clustering using NLP Techniques." Medium*, July 2023. https://medium.com/@danielafrimi/text-clustering-using-nlp-techniques-c2e6b08b6e95.
2. Bart, Yakov, & Kronrod, Ann. *"The Paradox of Language Repetition in Product Reviews." SSRN Electronic Journal*, 2018. https://doi.org/10.2139/ssrn.3194326.
3. Learn by Building AI. *"A Beginner's Guide to Building a Retrieval Augmented Generation (RAG) Application from Scratch." Learn by Building AI*. Accessed February 3, 2025. https://learnbybuilding.ai/tutorials/rag-from-scratch.
4. Weatherbed, Jess. *"Yelp's New AI-Powered Review Filters Will Show What You Want to Know." The Verge*, December 10, 2024. https://www.theverge.com/2024/12/10/24317749/yelp-ai-powered-review-insights-filters-availability.
5. Mamata Das, Selvakumar K., P.J.A. Alphonse. *"A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset."* Accessed Aug 8 2023. https://arxiv.org/abs/2308.04037

02 Generative AI statement

We used ChatGPT in the process of drafting this proposal for three main purposes:

- Ensure proper formatting of citations
- Confirm the feasibility of plans for the project
- Identify grammatical errors and improve sentence fluency

The links for the conversations relevant to the prompt are included below:

1. https://chatgpt.com/share/67a15c04-bc70-800d-94f4-1e6019e5a55b
2. https://chatgpt.com/share/67a15c22-67cc-800d-b11f-35e7cf0e7ce8