DATA 228: Big Data Technologies and Applications
Section 21

Project Proposal

Customer Segmentation And Analysis with Yelp Dataset

Group 6:
Sakshi Manish Mukkirwar (016794765)
Keerthana Raskatla (016780855)
Swati (016702413)
Akanksha Tyagi (016738839)
Anjali Himanshu Ojha (016803033)

Customer Segmentation And Analysis with Yelp Dataset

1. Abstract

In today's data-driven world, restaurants strive to acquire a competitive advantage by analyzing diner preferences and behavior. The Yelp restaurant reviews dataset, which contains millions of diner reviews, ratings, and restaurant profiles, is a significant resource for such customer insights. This research proposal highlights our intention to use Big Data analytics to do consumer segmentation utilizing this rich Yelp dataset.

As the size of data grows, identifying a segment with certain attributes is a core of any big-data system. As part of our Big-Data Applications project we are going to explore Yelp Restaurant Reviews dataset. As the data size keeps increasing, querying a large amount of data becomes very time consuming and computation expensive. For any modern marketing system having a great segmentation engine increases the time to market.

2. Motivation

The motivation for this research stems from the growing importance of data-driven decision-making in the business world. The Yelp dataset is a comprehensive collection of data from the Yelp platform, providing valuable insights into businesses, user reviews, and user profiles. It includes details about businesses such as their names, locations, attributes, and categories, along with user reviews containing star ratings and text. User profiles offer information about reviewers, including their names, review counts, and voting statistics. The dataset also covers check-in records, tips, and photos related to businesses. With access to this rich dataset, businesses have an unprecedented opportunity to harness data for understanding customer behavior and preferences.

Customer segmentation is a crucial strategy for personalized marketing, and by understanding the diverse customer groups that frequent businesses on Yelp, companies can optimize their offerings, improve customer experiences, and drive revenue growth. This research aims to bridge the gap between data analysis and actionable insights for businesses operating in the digital age.

3. Literature Survey

The phenomenal rise of online platforms and user-generated content in recent years has given researchers and businesses unparalleled access to valuable data sources. Among these, the Yelp dataset, a collection of restaurant reviews and ratings, has received a lot of interest in the Big Data analytics world. This literature review investigates previous research and studies on the Yelp dataset, with an emphasis on its applicability in customer segmentation, which serves as the foundation for our proposed project.

Luo et al. (2020)[1] build optimal recommender systems to address the issue of information overload on review websites. Their goal is to forecast the association between a reviewer's assessment of individual restaurant features (importance and sentiment) and overall satisfaction (number rating). They find five significant features using a modified Latent Aspect Rating Analysis technique. Notably, they believe that "restaurant value" is extremely crucial from the user's standpoint, whereas "food & drinks" has a substantial influence on sentiment. Furthermore, "restaurant value" is identified as the most important contributor to satisfaction. They advocate for the incorporation of "dynamic" recommender systems that take attribute-specific evaluations into account in order to enable enhanced, personalized navigation

of rich review information. This study provides a useful approach for improving the analysis of comprehensive reviews.

Kwon et al. (2021)[2] explore big data analytics by leveraging a large dataset of 1.5 million Yelp restaurant reviews. They create models to predict review helpfulness using machine learning approaches specialized for massive data, such as XGBoost. Key findings show that reviewer credibility influences reader perception more than numerical ratings or substance. This study demonstrates how big data analytics may be used to extract insights from large-scale online review databases. The findings have practical significance for hospitality firms who want to use big data to improve consumer decision-making by emphasizing influential review qualities.

Meek et al. (2021)[3] study the contextual and descriptive features of online restaurant reviews that influence users' perceptions of them as useful. They use qualitative and quantitative methodologies to study a large dataset of 58,468 Zomato evaluations, drawing on Dual Process Theory and Social Impact Theory. The main findings highlight the importance of good framing, solid argument quality, and moderate ratings in evoking "Likes" from readers. The study sheds light on the critical function of heuristics in filtering and amplifying the social impact of evaluations, providing potential consumers with an informed decision-making tool while boosting the platform's value. This big data exploration sheds light on the informational and normative elements that influence review usefulness.

Moon et al. (2021)[4] develop a novel market segmentation methodology that makes use of online consumer reviews to profile both customers and businesses on social media. Their two-sided technique divides reviewers based on declared preferences and companies based on

their reviewed practices after analyzing a huge Yelp restaurant review dataset. As a result, a comprehensive profiling of reviewer and business sectors is produced, providing actionable insights for targeted segmentation efforts on social media. This novel method goes beyond typical segmentation methods based on surveys or transactions. The research highlights the power of this methodology for firms seeking granular segmentation insights from social media data by leveraging publicly available, thorough consumption details inside online reviews.

4. Methodology - experiment design, algorithms to be used, evaluation methods

Experiment Design

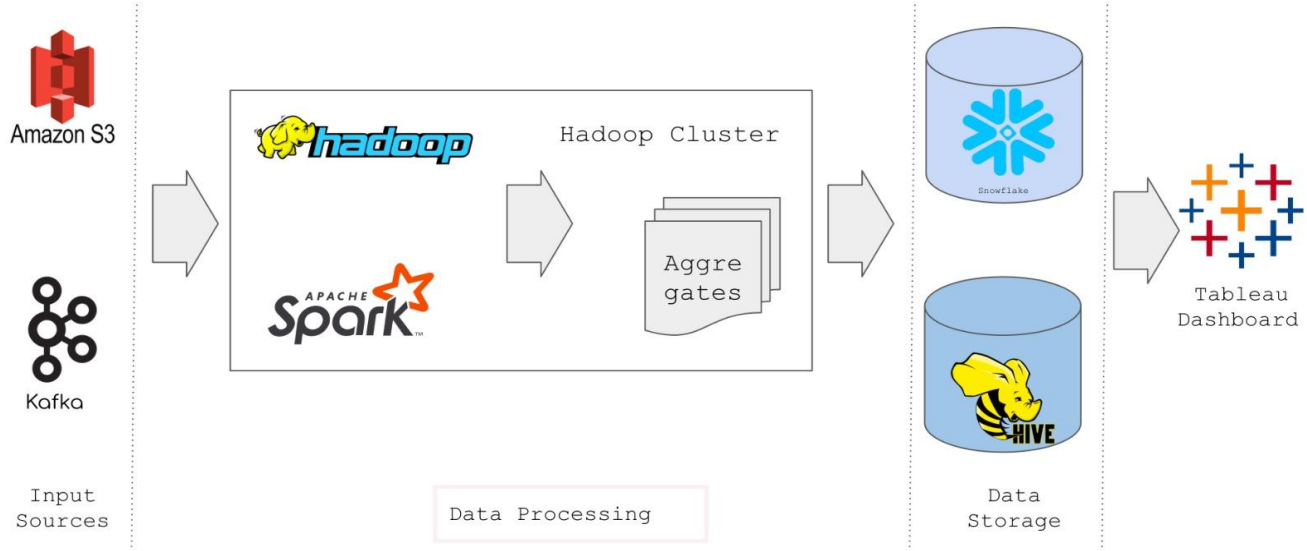

Fig 1. Overview of the System Design

***Data Collection*:** By collecting the Yelp dataset, ensuring we have access to various data files, such as business information, user reviews, and user profiles. We will take the dataset from the Yelp website (https://www.yelp.com/dataset) in which each file is composed of a single object type, one JSON-object per-line. Data has **6,990,280 reviews**

for **150,346 restaurants** in **11 metropolitan areas** given by **1,987,897 customers.** It also contains information about cuisine served in the restaurant, ambience and other information. Data size is 11.8 GB in compressed (.tar) format. Data and Attribute information -

Entity	**Fields**
Business	business_id, name, address, city, state, postal_code, latitude, longitude, stars, review_count, is_open, attributes, categories, hours
User	user_id, name, review_count, yelping_since, useful, funny, cool, elite, friends, fans, average_stars, compliment_hot, compliment_more, compliment_profile, compliment_cute, compliment_list, compliment_note, compliment_plain, compliment_cool, compliment_funny, compliment_writer, compliment_photos
Review	review_id, user_id, business_id, stars, useful, funny, cool, text, date
CheckIn	business_id, date
Tips	user_id, business_id, text, date, compliment_count

***Data Preprocessing*:** Cleaning and preprocessing the data, handling missing values, and ensuring consistency in data formats. This includes merging relevant datasets and encoding categorical variables.

***Feature Engineering*:** Extract relevant features and attributes for customer segmentation, such as user review patterns, sentiment scores, and user demographics, what kind of food they like, how many times they visit a restaurant, typical rating etc. Create feature vectors that represent users.

***Exploratory Data Analysis (EDA)*:** Perform EDA to gain insights into the data's distribution, correlations, and outliers. EDA will guide the choice of features and clustering algorithms.

Algorithms to be used

Clustering: K-means clustering can be used for market segmentation. In this algorithm, we cluster the data into K groups. This allows us to segment groups of people that have the same preferences.

Multi-dimensional visualization: We'll use visualization tools provided by tableau to create multi-dimensional visualizations. Such visualizations can help in showing us the distribution of users' preferences. We'll also use pairwise visualization tools to see the relationships between different features.

Dimensionality Reduction: Dimensionality techniques like PCA can be used to reduce the number of dimensions. This can help in improving the segmentation.

Evaluation Methods

End-to-end system testing: The most important evaluation of our system is an end-to-end test. We'll perform a comprehensive end-to-end test to make sure that each component of our system is performing as expected. This involves data loading, processing, storing and visualization modules.

Segmentation evaluation: We are creating a data cube, consisting of all the information about customers and their behavior. Given the large size of data, the data model will make the execution of the query very fast allowing us to explore data more quickly.

5. Deliverables - Including the possibility of writing a technical paper to submit to a journal, etc., and milestones.

Deliverable Description	Details	Delivery Date
Project Proposal	Project Proposal with data selection and high level idea about what we want to implement and achieve using Big Data tools and technology.	2023-10-06
Gathering and Refining Data	Prepare the data cleaning and formatting scripts.	2023-10-15
Exploratory Data Analysis	Start investigating data to identify useful attributes and how each attribute can be leveraged for the targeting. Write spark jobs to generate aggregates.	2023-10-26
Mid Project Review	Project Intermediate Status Report	2023-10-27
Cluster And Environment Setup	Start setting up cloud infrastructure where we will host our big data application	2023-11-05
Visualizations With Tableau	Once data is processed and saved in aggregates format, we will start creating a tableau dashboard with all useful information.	2023-11-15
Final Report and Presentation Delivery	Final report writing and presentation preparation. Compile all the useful insights which will be helpful in customer segmentation.	2023-12-01

6. Team members and their roles - Is the workload uniformly distributed?

Name	Role and Responsibilities
Akanksha Tyagi	Exploratory data analysis with Spark.
Sakshi Manish Mukkirwar	Visualization with Tableau.
Swati	Investigation of the data to identify useful attributes for targets.
Keerthana Raskatla	Gathering and refining data.
Anjali Himanshu Ojha	Cluster and environment setup.

The above table shows the direct responsibility of each individual. However, all the members of the team will help each other. Everyone will contribute to the final report as well.

7. Relevance to the course - Scope of the project falls within the topics covered in the course?

In this course we will use the Big-Data Tech stack taught in the class. For data processing we will use it for faster performance, Hadoop will be used for data storage, Kafka will be used for data consumption and a data store will be used for final aggregates. Later these aggregates will be used for data visualization. All the intermediate datasets will be stored in parquet format in hadoop for faster retrieval.

8. Technical Difficulty

One potential technical challenge is effectively managing large-scale, heterogeneous data from Yelp. The dataset contains a wide variety of data types, including text reviews, geolocation

information, and user profiles, all of which may necessitate sophisticated preparation and computational resources. Furthermore, assuring scalability and efficiency in applying advanced machine learning algorithms for consumer segmentation and analysis on such a heterogeneous dataset can present processing speed and memory utilization difficulties.

9. Novelty - Uniqueness

This dataset provided by Yelp is real data coming from real business and reviews. Our approach for this will not just be restricted to analysis, we are going to explore this data keeping restaurants in mind. Using this project we want to help restaurants find an audience for targeting. By narrowing our focus to restaurants, we acknowledge that they face distinct challenges and opportunities within the realm of online reviews and customer engagement. We will have all data related to customers like demographics, behavior attributes etc, in a form of data cubes. Unlike regular analysis, our project tailors its findings to address the practical needs of restaurants, offering them tangible strategies for audience targeting.

10. Impact - Chances of publication and utility

Effective targeting can be a very powerful tool to directly reach out to your target audience. Helping restaurants in understanding their customers can have a huge impact on their business. There are many aspects in this data which can be leveraged to quantify the customers behavior and expectations. Having all the data at one place with derived insights has a huge potential for businesses.

References

[1] Luo, Y., Tang, L. (Rebecca), Kim, E., & Wang, X. (2020). Finding the reviews on Yelp that actually matter to me: Innovative approach of improving recommender systems.

[2] Lee, M., Kwon, W., & Back, K.-J. (2021). Artificial intelligence for hospitality big data analytics: developing a prediction model of restaurant review helpfulness for customer decision-making.

[3] Meek, S., Wilk, V., & Lambert, C. (2021). A big data exploration of the informational and normative influences on the helpfulness of online restaurant reviews.

[4] Moon, S., Jalali, N., & Erevelles, S. (2021). Segmentation of both reviewers and businesses on social media.