Yelp Data Analysis using Hive

Philip Wong, Pratiksha Yadav, Pooja Madhup, Shailja Pandit
Department of Information Systems, California State University
Los Angeles
Tel. 323-343-2916, Fax. 323-343--5209
e-mail : pwong4@calstatela.edu, pyadav@calstatela.edu, pmadhup@calstatela.edu, spandit3@calstatela.edu

Abstract: The goal of this project was to provide a high-level, general, and aggregated approach to sentiment analysis. We will identify the overall sentiment of yelp users (also known as yelpers) and key behaviors. In our analysis, we focus on two categories: healthcare and food/beverage. Questions we hope to answer are: What insights can we obtain from the US healthcare system? What does the data show about the food establishments across the US?

1. Introduction

Yelp is an internet-based company that focuses on crowd-sourced reviews for millions of businesses and establishments across the US and globally. Yelp users can submit feedback in terms of reviews, ratings, and list information pertaining to the establishment.

The yelp data set will include three Json files (Business, User, and Reviews) which we will join with the dictionary dataset required for sentiment analysis.

2. Related Work

Generally, many works concentrate on using predictive models to determine user sentiment. In Sentiment Analysis of Yelp Data – Text classification the author attempted to use 2 training models to predict user sentiments. [1] The author concentrated on the quality of the model and n-gram analysis. It did not provide many business-related insights related to the sentiment of the users besides a general sentiment distribution. Our work differs as we applied the sentiment analysis data to a specific category for business insight, in addition, we also determined the degree of positive and negative a yelp user's polarity was using an aggregated-sum approach to the sentiment analysis.

A related study and analysis "Collecting and Analyzing Patient Experiences of Health Care From Social Media" [2] were performed in 2015 where they used sentimental analysis techniques and Hadoop to calculate the sentiment score of each sentence with parallel processing. In their study, they mainly focus on 26 healthcare-related categories (examples include hospitals, urgent care facilities, and medical centers) to extract healthcare-related businesses from the Yelp academic dataset. Whereas, we have used hive to manipulate the data and build queries which are then used to analyze data and get useful insights and visualizations. Our analysis was distributed across several categories like restaurants, hospitals, etc. Using this data analysis, businesses can improve their product/services and fine-tune their approach toward better overall service. Moreover, through a geospatial map, we also determined which part of the city has the greatest number of yelpers.

A paper by Zeynepderbent [3], 2021 attempts to come up with a Sentiment analysis on Yelp Restaurant Reviews. This project also focuses on predicting sentiment value for a given review by creating a classification model. They used python libraries to create the visualization, whereas in our project we used the Tableau tool for creating visualizations.

3. Dataset Info

- DATASET NAME: Yelp Dataset
- DATASET URL: https://www.yelp.com/dataset/download
- TOTAL SIZE: 8.21 GB
- COUNTRIES CONSIDERED: USA and World
- NUMBER OF FILES: 3
- FILE FORMAT: JSON, CSV

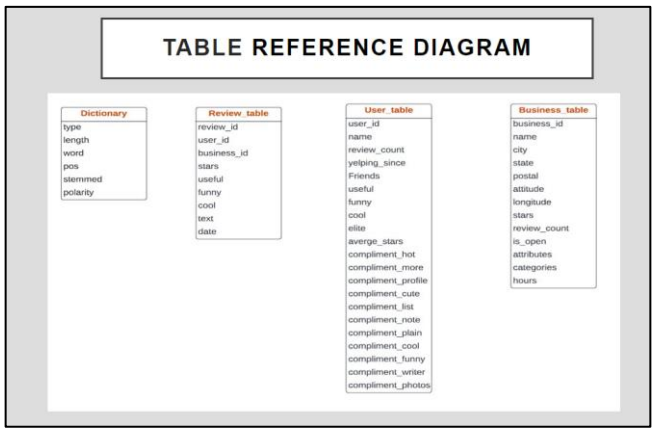

Fig 3 Table Reference Diagram

Fig 3 shows the different tables- Business, User, and Review we have used and the related columns to each table.

4. H/W Specifications

- VERSION: 3.2.1-amzn-3.1
- TOTAL NODES: 3
- TOTAL NODE MEMORY SIZE: 0-7 / 30874 KB
- TOTAL NUMBER OF CPUS: 4
- TOTAL STORAGE: 481GB

5. Background Work

Apache Hadoop is a collection of open-source software that consists of Linux-based structure tools and is used to run applications for big data. Big data consists of a massive amount of data that is in an unstructured format. It has 2 core components – Map reduces where the data can be processed and Hadoop Distributed File System (HDFS) to store the data. These consist of 2 nodes in HDFS – data nodes where the data is stored in replicated file blocks and the name node which forms a relationship between the client and the data node. The Apache Hive JDBC client such as beeline is the primary way to access Hive. Hive is an SQL-based data warehouse system where the queries can be performed and open the door for analysis of massive data set stored in HDFS.

6. Project Instructions

6.1 Workflow Architecture

The first step was to find out the right source of data. In this data set of yelp that we have chosen, we have defined objectives in terms of measurable goals we wanted to achieve by the end of the project. We downloaded the data from the yelp website. Using the SCP command we have uploaded data to HDFS. Using hiveql we extracted data from different files where we created queries based on the use case. Further for visualization we used tools like tableau and excel power map. Fig 6.1 shows the Workflow architecture.

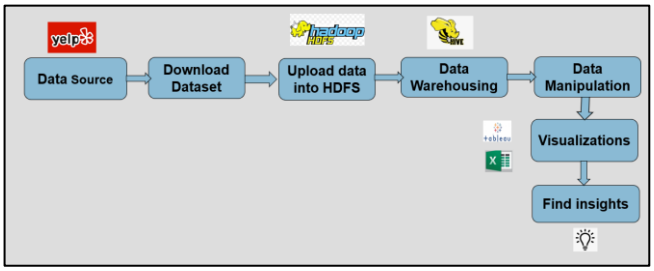

Fig 6.1 Workflow Archtecture

.

6.2 Data Cleaning

Generally, a minimal amount of manual cleaning was required for the dataset. We used specific filters and strategic joins to minimize any null and empty fields. We analyzed the output of each query at each step of the workflow process to determine whether the data was relevant and valuable in our use for visualization.

6.3 Visualizations and Insights

We have created all the Visualization using Tableau and excel tools. Below are a few of the Visualizations we have achieved as a part of this project.

6.3.1 Yelp Hospital sentiment

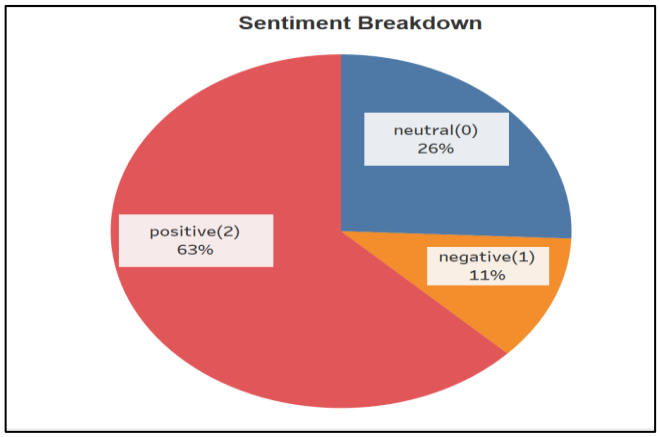

Fig 6.3.1 Yelp Hospital Analysis

The visualization in Fig 6.3.1 represents the sentiment breakdown of hospital categories for yelp users in the United States. Here, in the graph, we show that they are three types of sentiment that are positive, negative, and neutral and the number shows as per the sentiment type. We analyzed the pie chart using the Tableau tool. From the above analysis, we found that the positive sentiment of the hospital category for yelpers is 63% rounded off, while the negative sentiments are 11%, and the remaining 20% of yelpers have neutral sentiment for the hospital category. Therefore, we conclude that most of the yelpers have positive sentiment regarding the hospital category and they are satisfied by their hospital services.

6.3.2 Total Yelpers using Hospital as Category in Yelp for Different cities in the USA

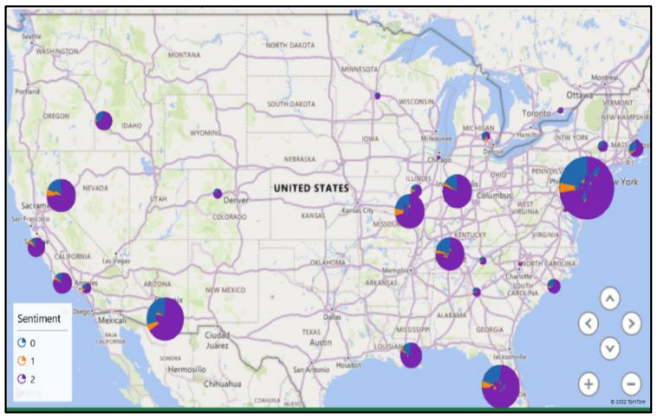

Fig 6.3.2a Yelper sentiment for Hospital

The visualization in Fig 6.3.2a shows the total yelpers using the hospital category in yelp for different cities in the USA with over 2 million population. Using the 3D power map tool, we analyze the positive sentiment, negative sentiment, and neutral sentiment which are numerically defined as 2,1,0 respectively. In the above geographical map, we analyze that most of the northeast region side of yelpers have a positive sentiment for the hospital category. From the above map, we also identified that most of the cities have positive sentiments throughout the United States. Therefore, we can say that the positive polarity was present overall with respect to the hospital category. In the second observation, we found that the yelp users of Philadelphia cities use yelp the most for the hospital category than any other city in the USA, however, California cities have the lowest yelp user using the hospital category. We found this city for the most users due to the limited dataset to be visualized in the 3D power-map and the limited dataset being offered by the yelp.

User Sentiment of Hospital for Pennsylvania State

The following visualization Fig 6.3.2b shows a drill-down of Pennsylvania state for reference. We can see the general sentiment for the hospital is overwhelmingly positive with few negative reviewers.

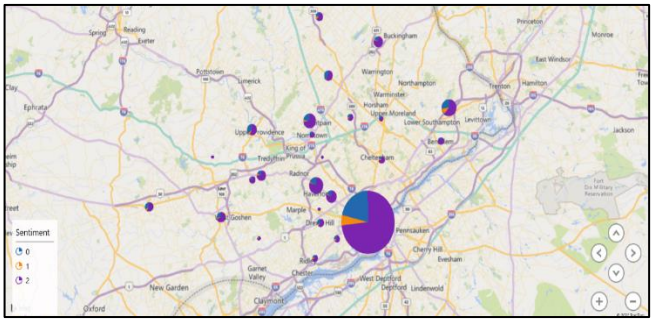

Fig 6.3.2b User sentiment of Restaurants for PA state

6.3.3 User Sentiments of Restaurants By City

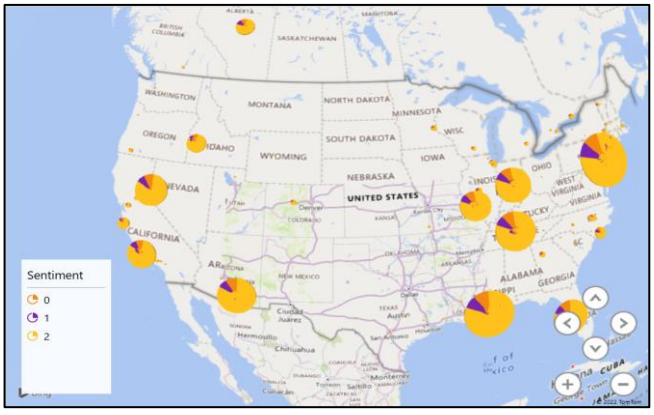

Fig 6.3.3a User Sentiments of Restaurant By City

In this analysis Fig 6.3.3, we visualized over 2million records group by the city. This analysis represented approximately 2/3 of the data set through the Excel 3D Powermap tool. The Sentiment legend depicted identifies Positive sentiment identified as 2, Negative sentiment is identified as 1, and finally Neutral sentiment is identified as 0. We identified that sentiment is very consistent throughout the country. Overall positive polarity was identified throughout the country in regard to the restaurant category. Secondly, we identified that most yelp reviewers reviewed in Pennsylvania than in any other city. The main reason for this is the limited dataset to be visualized in the 3d Power-map and the limited dataset being offered by yelp.

User Sentiment of Restaurants for Pennsylvania State

The following visualization Fig 6.3.3b is a drill-down of Pennsylvania state for reference. We can see the general sentiment for restaurants is overwhelmingly positive with few negative reviewers. Also, despite filtering the data we were able to produce more than one million reviews in Pennslyvania which still limited our visualization to approximately 800k records.

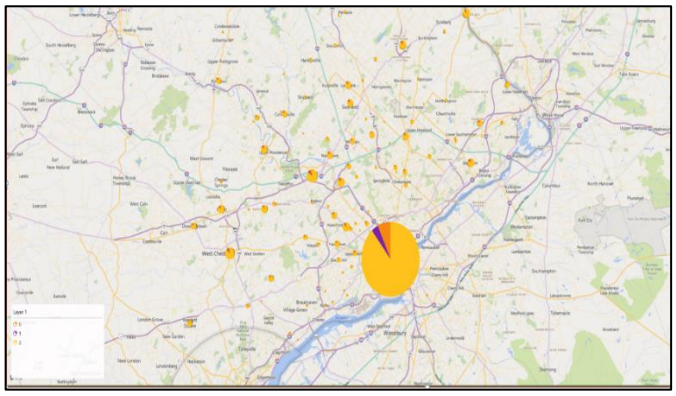

Fig 6.3.3b User Sentiment of Restaurants for PA state

6.3.4 Top 10 states with the highest review count.

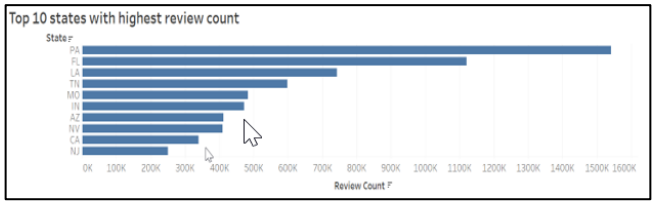

Fig 6.3.4 Top 10 states with the highest review count

This visualization Fig 6.3.4 represents the sum of the review count of the top 10 states in the Us. For this analysis, we used a sum of the review count and we grouped by a state to find the top 10 states. As we observe from the bar graph Pennsylvania has the highest review count followed by Florida, Louisiana & Tennessee. However, due to the data limitations in the dataset, we had biases in the review count for Pennsylvania.

6.3.5 Total count of yelp users for the last 10 years

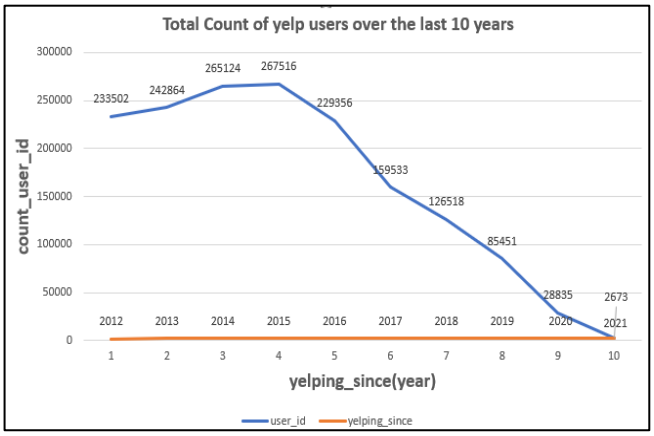

Fig 6.3.5 Total yelpers count for last 10 years

The visual data Fig 6.3.5, shows in which year yelp has got maximum users using the yelp app or yelp.com to post their reviews. Yelp got the maximum number of users in 2015 and from 2016 onwards the count of users using yelp is declining. The decline in user reviews could be caused because of the incompleteness of the datasets.

6.3.6 Aggregated Sum of Sentiment by User

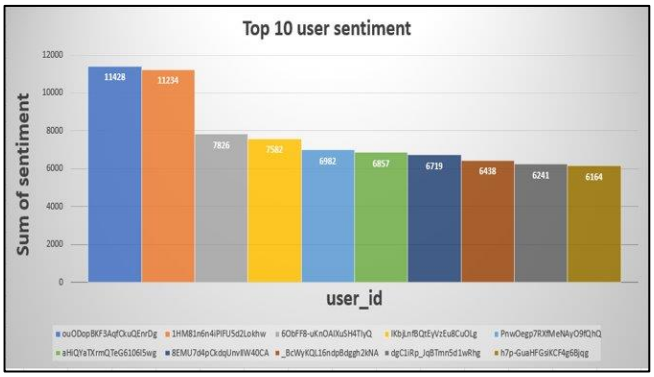

Fig 6.3.6a Top 10 user sentiments

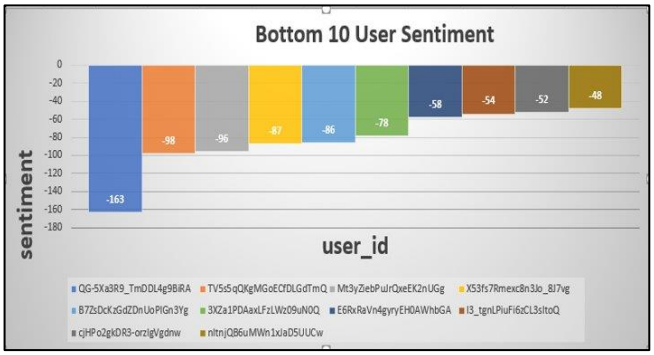

Fig 6.3.6b Bottom 10 user sentiments

The process we used to determine the aggregated-sum is a process of breaking down the reviews into words, then applying the polarity from the dictionary file, as referenced in the table reference diagram, to each word. Then we grouped each word polarity to each review. Finally, we grouped the reviews to each user and summed the quantitative polarity. The ultimate output we obtained was the "NET" polarity of each user which provided an overall character, whether positive, negative, or neutral for each user. We filtered for the top and bottom users. We found users were extremely positive when "NET" polarity was high – most likely caused by fake accounts used to provide positive reviews. Negative polarity users were what was expected. The average or mean polarity was positive 16, which contributed to the exceedingly positive nature of the top users. Refer Fig 6.3.6a, Fig 6.3.6b for the charts.

7. Conclusion

Yelp users are overall satisfied with US hospitals as a whole throughout the country. Similarly, yelp users have a positive sentiment toward food. Most yelp reviewers were reviewing establishments from Philadelphia which was due to the limitations of the dataset analyzed. Yelp users peaked in the year of 2015.

8. Github Link

https://github.com/pooja9050/Yelp-Data-Analysis-Using-Hive

9. References

[1] Abhijeet Singh, 'Sentiment Analysis of Yelp Data- Text Classification', Analytics Vidhya ,2021 Sentiment Analysis of Yelp Data — Text classification | by Abhijeetsingh | Analytics Vidhya | Medium

[2] Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S Collecting and Analyzing Patient Experiences of Health Care From Social Media, 2015
URL: https://www.researchprotocols.org/2015/3/e78

[3] Zeynepderbent, 'Performing Sentiment Analysis on Yelp Restaurant Reviews' , Medium , 2021
https://medium.com/analytics-vidhya/performing-sentiment-analysis-on-yelp-restaurant-reviews-962334d6336d

[4] Reference for Data dictionary table
https://github.com/dalgual/aidatasci/raw/master/data/bigdata/dictionary.tsv