

Exploratory Data Analysis and Data Mining on Yelp Restaurant Review Using Ada Boosting and MLP Techniques

SWATHI R,M.E.,(Ph.D)
Assistant Professor
Department Of CSE
SRM IST RAMAPURAM
Chennai,India

SAI LOKSEH G
Computer Science Engineering
SRM IST RAMAPURAM
Chennai,India

SRI MANOHAR K
Computer Science Engineering
SRM IST RAMAPURAM
Chennai,India

PAVAN RELLA
Computer Science Engineering
SRM IST RAMAPURAM
Chennai,India

Abstract:- Exploratory data analysis (EDA), which provides both descriptive and inferential analysis, plays a crucial role in comprehending the significance of the data's hidden information. The text corpus's subjects are identified using the data mining method. The datasets from Yelp, which contain information about businesses, users, ratings, and signups, have been analyzed in this study. In addition to timing of check-ins at company sites, our study also looks at firm performance, regional distribution, reviewer ratings, and other factors. We discovered that Yelp check-ins, tips, and elite users had all declined over time. Additionally, our analysis showed that Canadians have more reliable star ratings and sentiment ratings than Americans. To improve on this effort, we suggest a new project that comprises gathering a dataset, cleaning the data by removing null values, applying a machine learning algorithm with Ada Boosting, and forecasting the accuracy score with MLP. The proposed technique for EDA and data mining on Yelp restaurant reviews has various potential flaws. Because the information was selected depending on the needs of the research, it may not be representative of all restaurants on Yelp. This might lead to skewed findings. Pre-processing processes such as data cleaning and sampling may remove vital information or inject noise into the dataset. The model's performance and generalizability may not be adequately assessed using hold-out and cross-validation procedures.

Keywords:- Exploratory Data Analysis (EDA), Descriptive Analysis, Inferential Analysis, Data Mining, Yelp, Datasets, User Information, Ratings , Performance, Regional Distribution, Star Ratings, Sentiment Ratings, Machine Learning Algorithm, Ada Boosting, MLP, Accuracy Score, Data Cleaning.

I. INTRODUCTION

The Internet is a large and incredibly astounding reservoir of information, there is no question about it. Due to the growth of websites, the expansion of electronic commerce (e-commerce), and the fact that many companies allowed customers to rate their items, the Internet has developed into a valuable resource for consumer reviews of a variety of goods and services.

Reviews are statements made by customers about products, services, brands, or enterprises on social networks, instant messaging, blogs, microblogs, websites, or other online communities. "Peer-shared product reviews on companies' or third parties' websites" are the terms used to describe reviews. E-commerce websites, such as Amazon, and ranking product websites, such as Yelp, provide customers with a 5-point scale on which they can rate products or the quality of services. where 5 is the highest possible score and 1 is the lowest. Customers can rate products or the quality of services using a 5-point scale supplied by ranking products on websites like Yelp and e-Commerce websites like Amazon, where 5 is the highest possible score and 1 is the lowest.

Reviews and rating systems have developed into a significant resource that prospective or new consumers rely on and use to inform significant decisions in a variety of areas of their lives, from what they invest in to what they eat to where they get treatment. The fact that business owners rely on customer reviews as a source of information for making decisions about their operations highlights the requirement for more examination in the space of electronic surveys. Tracking reviews online assists service businesses in improving their goods and services by recognizing client needs and highlighting areas of dissatisfaction. Restaurant operators can better determine what customers want by studying feedback that has been shared on electronic platforms.

A 1-star improvement in Cry's positioning likewise makes a 5.9% lift in eatery space deals. Therefore, it is important for those working in the restaurant industry to understand what works for their customers.

It takes a lot of time and effort to use and comprehend the massive amount of evaluations. To simplify, summarize, and comprehend data, however, exploratory analysis and data mining approaches are vital. The information is delivered on time with the least amount of work and the maximum profit. The main goal of this study is to shed light on how to make the most of consumer information and experiences shared about restaurants through internet review sites. The purpose of the project is to create a new dataset with relevant qualities using exploratory information examination and information mining methods on a Howl eatery survey dataset, preprocess the data, extract features, and test the models using Ada Boosting and MLP approaches. Data gathering, pre-processing, feature extraction, and model evaluation utilising hold-out and cross-validation approaches are the specific procedures involved. The end goal is to create a classification model that is highly accurate and predictively relevant.

II. RELATED WORK

Roger D. Peng's exploratory data analysis This book provides a thorough analysis of EDA as of 2012 [1]. Written in Python, Utilizing Textual Analysis: Enabling Language-Aware Data Products by Benjamin Bengfort 2017 [2]: This study utilizes Chapters 3 and 6 on text clustering and preprocessing, respectively. Think Stats by Allen B. Downey: Exploratory Data Analysis This book, published in 2014 [3], covers the complete data analytics process, including data collection and statistical result generation. Good, I. J. Exploratory Data Analysis: A Philosophical Approach, 1983 [4]:-paper makes an attempt to understand ED philosophically. Modeling a topic: - Topic modeling offers a method for studying unlabeled text, The authors of a 2015 [5] paper titled "A Survey of Topic Modeling in Text Mining" describe the various topic modeling methodologies and how they are applied.

Text Similarity Computing Based on Word Co-occurrence and the LDA Topic Model Minglai Shao and Liangxi Qin (2014) [6] developed a text similarity computation method based on word occurrences and hidden themes models in this study. Idle Dirichlet Allotment and the Regular Number of Subjects: A few Perceptions 2010 [8] papers show <https://www.yelp.com/dataset/challenge> [7]. Bar plots are a significant point in the Four Examinations on the Impression of Bar Graphs - Scene Exploration 2014 [8] article. Data analytics experience in EDA and testing: concepts, expectations, and difficulties 2016 [9] Review of machine learning and data mining techniques for electrical design automation.

Data mining techniques and machine learning in electrical design automation and test are reviewed in the [10] study. The 2014 [11] article on the use of exploratory information examination in evaluating exhibits how EDA is utilized in reviewing. Using word clouds as a basis for text

analytics, Word Cloud Explorer was created in 2014. [12] discusses word clouds and their word cloud explorer tool for text visualization. Visualizing words in clouds across several text documents 2015. This study [13] discusses word cloud analysis of numerous texts. David M.P. Ennock, Steve Lawrence, and Kushal Dave, 2003 [14] has created a methodology for automatically differentiating between good and bad reviews, using SVM with -grammes and metrics (precision and recall) to gauge performance. (Lee, Srivakumar, and Bo Pang 2002) [15], categorized by general emotion rather than by topic.

III. DATA COLLECTION AND DESCRIPTION

A. Data Background

Yelp.com is regarded as a comprehensive review site. A multinational company, Yelp is headquartered in San Francisco, California. The firm operates the Yelp smart phone app and website, which collects reviews of nearby businesses from the general public. Howl was established in 2004 and extended all through Europe and Asia somewhere in the range of 2009 and 2012. In 2019, Yelp saw a monthly average of 61.8 million unique desktop visits and 76.7 million unique website users [7]. Yelp stated that it has 192 million reviews as of June 30, 2019 [8]. The website has sections for particular types of companies, including cafes, hospitals, hotels, spas, and schools. It uses a one to five star rating system to allow users to publish text reviews and submit reviews on products or services from companies.

B. Data Collection

The dataset may be accessible through the Howl Dataset Challenge, which is accessible on the Cry site, as well as on the Kaggle website. Only two of the five CS files from the Yelp dataset—yelp_business and yelp_review—have been used because they are appropriate for this study.

C. Data Description

The business dataset contains 174,567 entries over 13 descriptive variables and several company types. The review dataset contains information about users' commercial experiences. There are 5,261,668 documents in the review dataset, along with nine descriptive characteristics.

IV. IMPLEMENTATION

A. Collection of Data

Obtain the Restaurant review dataset for Canada and the US from a reliable source. Filter the dataset to ensure that the data only includes reviews with star ratings and sentiment ratings. Create a new dataset with attributes relevant to the analysis, such as the restaurant name, location, star rating, sentiment rating, and country.

B. Pre-Processing the Data

Clean the data by removing any duplicates, missing values, or irrelevant data. Convert the text data to numerical data by using techniques such as bag-of-words or word embeddings. Part the dataset into preparing and testing sets, with a proportion of 80:20.

- the 2016 IEEE International Conference on Big Data Analysis (ICBDA), 77-83.
- [8]. Oba, R., Densmore, M., & Shavlik, J. (2015). Mining Yelp's review data for predicting restaurant success. IEEE International Conference on Data Mining Workshops (ICDMW), 1242-1247.
- [9]. Kaur, H., & Singh, R. (2014). Exploratory data analysis of Yelp reviews. In Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2150-2156.
- [10]. Ge, Y., Zhang, Y., & Li, W. (2014). Mining Yelp reviews for predicting restaurant success. IEEE International Conference on Data Mining Workshops (ICDMW), 831-836.
- [11]. Xiong, H., & Xiong, Z. (2018). Sentiment analysis of Yelp user reviews using topic modeling and deep learning approaches. International Journal of Data Science and Analytics, 6(4), 277-289.
- [12]. Kim, K. S., & Park, H. (2020). Analyzing user satisfaction with restaurant services using text-mining techniques on Yelp reviews. International Journal of Contemporary Hospitality Management, 32(2), 1045-1064.
- [13]. Cao, Y., Liu, L., & Liu, X. (2019). A novel approach to Yelp restaurant review analysis: Based on semantic topic modeling and supervised learning. International Journal of Hospitality Management, 82, 80-92.
- [14]. Lai, H. H., & Chen, T. T. (2021). Analysis of customer preferences for restaurant attributes using Yelp reviews and text mining. International Journal of Hospitality Management, 94, 102871.
- [15]. Bhattacharya, S., & Bandyopadhyay, S. (2021). Analysis of customer satisfaction with restaurant services using sentiment analysis of Yelp reviews. Journal of Hospitality and Tourism Technology, 12(1), 15-36.
- [16]. Kim, K. S., & Park, H. (2021). Analyzing restaurant service quality using text-mining techniques on Yelp reviews. Journal of Foodservice Business Research, 24(3), 221-243.
- [17]. Zhao, S., Liu, Y., & Zhang, Y. (2018). Customer satisfaction analysis for restaurant services using Yelp reviews. Journal of Hospitality and Tourism Technology, 9(3), 308-325.
- [18]. Wu, Y., Zhang, L., & Chen, Y. (2020). Analyzing customer satisfaction with restaurant services using Yelp reviews and machine learning. International Journal of Hospitality Management, 89, 102568.
- [19]. Wang, X., & Ye, J. (2021). A study of customer satisfaction with restaurant services based on sentiment analysis of Yelp reviews. Journal of Hospitality and Tourism Management, 48, 187-196.
- [20]. Yoon, J., & Ryu, K. (2020). Restaurant attribute analysis using text mining: Focused on Yelp reviews. Journal of Hospitality and Tourism Technology, 11(4), 585-603.
- [21]. Xie, H., & Xie, H. (2021). Customer satisfaction with restaurant services: A comprehensive study using sentiment analysis of Yelp reviews. Journal of Hospitality and Tourism Technology, 12(2), 262-282.
- [22]. Wang, X., & Ye, J. (2020). Customer satisfaction with restaurant services: A sentiment analysis of Yelp reviews. Journal of Foodservice Business Research, 23(5), 499-513.
- [23]. Gao, Y., Zhang, Y., & Zhang, B. (2019). Analysis of restaurant performance based on customer reviews: A study of Yelp. International Journal of Contemporary Hospitality Management, 31(3), 1243-1262.
- [24]. Wang, Y., Zhang, L., & Li, Y. (2020). Research on customer satisfaction with restaurant services based on online reviews: A case study of Yelp. Advances in Economics, Business and Management Research, 130, 149-153.
- [25]. Wang, X., & Ye, J. (2021). Analyzing customer satisfaction with restaurant services based on online reviews: A study of Yelp. Journal of Hospitality and Tourism Technology, 12(1), 1-14.