ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23476

Identifying Key Features in Yelp Data for Success in Different Types of Restaurants

Received:
21 August 2024
Accepted:
26 January 2025
Published:
23 February 2025

[1*]Andrianshah Priyadi, [2]Nelly Malik Lande, [3]Anita Faradilla, [4]Ma'arif Hasan, [5]Evi Widianti
[1-5]Researcher, National Research and Innovation Agency, Indonesia
E-mail: [1]andr028@brin.co.id, [2]nell001@brin.co.id, [3]anit010@brin.co.id , [4]maar001@brin.co.id, [5]eviw001@brin.co.id
*Corresponding Author

Abstract— Background: The purpose of this research is to measure of customer satisfaction for newly established independent restaurants and, consequently, good predictors of independent restaurant success. Urban communities face several challenges, including how to best use scarce resources like real estate and support small enterprises. Smart businesses are essential to the development of smart cities because they use data analytics to inform their strategic planning and design choices, and the target of this topic is restaurant. **Objective**: Restaurants control a sizable portion of the city market's small business sector. As part of the Yelp Data Challenge, Yelp just made available an open dataset that includes important details, ratings, and Yelp scores for every restaurant in different cities. **Methods**: Our methodology utilizes a vector of crucial factors to accurately forecast a business's prospective success and exclusively evaluate eateries located inside the city limits of Las Vegas. The dependent variables will consist of the mean Yelp ratings for each restaurant and constructed our model by following the subsequent stages. **Conclusion**: The findings of this research is corroborated by the discovery that the statistically significant properties of restaurants, shown by a low p-value, varied across various restaurant categories, the unique modeling technique to forecast future restaurants' Yelp rankings based on their design choices. This will assist owners of restaurants in making better design choices, which will result in more prosperous small enterprises in urban settings.

Keywords— Yelp Dataset Challenge; Linear Square Regression; Restaurant Attributes

Corresponding Author:

Andrianshah Priyadi,
Research Center for Energy Conversion and Conservation,
National Research and Innovation Agency,
Email: andr028@brin.go.id,
Orchid ID: https://orcid.org/0009-0005-8037-1913

I. INTRODUCTION

We are expecting a huge population increase in urban regions in the coming decade. The main reasons for this change could include job opportunities and potentially improved living standards. As the utilization of resources increases, so will the need for employment due to urbanization. The number of restaurants in cities will certainly need to increase accordingly to reflect this increase in population. Reflecting on the size of the industry, in 2016, the National Restaurant Association (NRA) found that restaurants would remain the second largest private sector employer in the United States, with a workforce of 14.4 million, $783 billion in sales, and its 7th consecutive year of growth [2].

Assessing the potential success of a certain restaurant can be challenging. Customer happiness is a crucial factor in determining the success or failure of a firm, as it directly impacts its financial performance. There are other factors that can contribute to a restaurant's failure to meet customer satisfaction, including choosing an unfavorable site or neglecting to provide essential services. Our objective is to address the issue of urbanization by offering accurate projections for aspiring restaurant proprietors. If a prospective proprietor possesses a well-defined blueprint encompassing the restaurant's type, location, and a comprehensive list of distinctive attributes, employing a model could facilitate the anticipation of the restaurant's success by drawing upon the accomplishments of analogous establishments within the vicinity [6],[11]. This information can assist prospective small business owners in determining the optimal location for their restaurants, determining the most suitable operating hours, and selecting the appropriate range of services to offer. As a result, there is an increase in the number of thriving small enterprises and a more optimal utilization of urban areas.

Yelp allows customers to rate businesses they have used on a scale of 1 to 5 stars for the benefit of other potential customers. It has been found through previous research that a 1-star increase in Yelp scores translates to a 5–9% increase in restaurant revenue [7]. This correlation was driven by independently owned restaurants, with chains playing a less prominent role. Therefore, it can be inferred that Yelp scores are a useful measure of customer satisfaction for newly established independent restaurants and, consequently, good predictors of independent restaurant success. The Yelp Data Challenge [1] contains business attributes such as location, services, and features, along with Yelp scores for 86,000 businesses in 10 cities and 4 different countries. The goal of this work is to find a novel modeling process that can predict the Yelp score of a potential restaurant, given the relevant features of that restaurant.

A consensus exists that restaurant success is most accurately predicted by its attributes, including but not limited to cuisine quality, supplementary services offered, ambiance, price range, noise level, and parking [3], [7]. This is even more true than the influence of location. In addition, among the numerous data analytic techniques that could be employed to ascertain critical attributes that contribute to customer satisfaction and experiences, Linear Least Squares Regression has been determined to be the most effective approach for comprehending the correlation between restaurant rating and the attributes [3]. Conversely, prior models attempting to forecast Yelp scores have proven to be unsatisfactory in terms of producing precise predictions [3].

On the premise that existing restaurants with comparable profits will serve as the most accurate predictors for a given potential restaurant, our model refines the outcomes discovered in prior research. To examine this hypothesis, we have applied ethnicity-based filters to the Yelp dataset pertaining to restaurants in the city of Las Vegas (Mexican, American, etc.). We subsequently employed four distinct methodologies, one of which was linear regression, to generate more accurate forecasts regarding the prospective success of the restaurant [4]. To validate our model, we generated forecasts for establishments that are already present in the city. Our investigation revealed that the accuracy of least-squares predictions can be substantially enhanced by employing comparable establishments. In addition, depending on the type of restaurant, the essential features for making these predictions may differ (for instance, distinct features are utilized to predict Italian restaurants as opposed to American restaurants). We hypothesize that our model could produce comparable results if applied to restaurants of any type in any city, even though its application was limited to four distinct ethnic restaurants in Denver. We also believe that the customer satisfaction of other categories of businesses could potentially be predicted using a comparable modeling procedure [12].

II. RESEARCH METHOD

Our methodology utilizes a vector of crucial factors to accurately forecast a business's prospective success. We exclusively evaluate eateries located inside the city limits of Las Vegas. In addition, we are developing distinct models for restaurants belonging to various categories. As an illustration, we will create a distinct model for Mexican restaurants, another for American eateries, etc. Thus, we consider the independent variables solely pertinent features, such as location, pricing, environment, and other services provided. We are utilizing the Yelp score of the restaurant as the criterion for measuring success. Consequently, the dependent variables will consist of the mean Yelp ratings for each restaurant [13]. We constructed our model by following the subsequent stages. The text processing was conducted using Python 2.7, whereas the modeling was performed using Matlab 2015. This phase involves an arithmetic procedure to determine the

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23476

similarity value between items following the conversion of the preceding dataset into a sparse matrix. At this stage, researchers assess the similarity value between products and the similarity value between users by employing the cosine similarity formula for both categories. Table 1 summarizes the studies on predicting review helpfulness on yelp business review sites.

Table 1. Summary of literature on predicting the helpfulness of reviews

Author Year	Preconditions for Review Helpfulness	Number of Review	Targeted Location	Methods	Main Conclusion
Kwok and Xie [8]	Word count; sentence count; gender of reviewer; age of reviewer; ratings; reviewer experience (status; membership; visited city)	56,284	Houston, San Antonio	Linear Regression	The efficacy of online hotel reviews is enhanced by managerial responses and the status of the reviewers.
Hu et al [9]	Evaluate quality; assess sentiment; analyse reviewer attributes	1,434,005	Chicago, Miami	Linear Regression, random forest	The review rating and word count predict the helpfulness of reviews across various users' travel regions, seasons, and sorts.
Li et al [10]	Temporal indicators (time-related terminology); explanatory indicators (causation-related terminology); sensory indicators (visual, auditory, tactile)	186, 714	Las Vegas	Negative binomial regresion	Temporal indicators exert the most significant influence on the utility of reviews.

The dataset utilized in this research is the Yelp Challenge Dataset, an openly available dataset comprising 2.7 million reviews, 556,000 business parameters, and the average Yelp ratings for 86,000 businesses. The firms are situated in several nations and cities: Edinburgh, U.K.; Karlsruhe, Germany; Montreal and Waterloo, Canada; Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, and Madison, U.S. Yelp, Inc. provided the dataset for the Yelp Data Analytics Challenge in 2016 [1]. The assessments we conduct are derived explicitly from restaurant data obtained from the city of Las Vegas. We successfully obtained data for approximately 4,800 eateries located in Las Vegas. The dataset is provided in JSON format, as depicted in Figure 1. When training and testing our model, we consider key features.

1. Cuisine type (mexican, american, chinese, etc.)
2. Attribute profile (including price range, availability of alcohol, etc.)

3. Geographical area (downtown, The strip, chinatown, etc.)
4. Mean Yelp rating (rated on a scale of 1 to 5 stars)

The primary modeling methodology applied in this paper is Least Squares modeling. Least Squares models are linear models that generate predictions for future data points according to the solution to the least squares problem using existing data:

$$\| \mathrm{A}\hat{\mathrm{X}} = \mathrm{b} \|_2 = \min_{x \in \mathbb{R}} \| \mathrm{AX} = \mathrm{b} \|_2 \qquad (1)$$

In the context of predicting Yelp scores, **A** is a matrix of the relevant attributes (columns) for each restaurant in the dataset (rows), **b** is a vector of Yelp scores for those restaurants, and the solution $\hat{\mathbf{x}}$ is a vector of computed correlation coefficients for each attribute. To make future predictions based on the data, $\hat{\mathbf{x}}^{\boldsymbol{T}}\mathbf{y}$ is multiplied by a vector of features for the new restaurant **y**, and $\hat{\mathbf{x}}^{\boldsymbol{T}}\mathbf{y}$ gives the Least Square prediction for the Yelp score of that restaurant.
The method we use to calculate the solution $\hat{\mathbf{x}}$ is the Moore-Penrose pseudoinverse. Let $\mathbf{A} = \mathbf{AU}\sum\mathbf{V}^{\boldsymbol{T}}$ be the singular-value decomposition of **A** where **U** and **V** are orthonormal square matrices and **Σ** = *diag*(σ1, ..., σr, 0, ..., 0), where ***r*** is the rank of $\mathbf{A}$. Then, the pseudoinverse is given by:

$$\mathbf{A}^{\dagger} = \mathbf{V}\textstyle\sum^{\dagger}\mathbf{U}_{\boldsymbol{T}} \qquad (2)$$

Where

$$\textstyle\sum^{\dagger} = diag\left(\frac{1}{\sigma 1}, \dots, \frac{1}{\sigma r}, 0, \dots, 0\right)$$

The advantage of this method is that even if A is rank-deficient that the pseudoinverse will still exist and may be utilized to calculate a minimum norm solution to the least squares problem [22]. In this scenario, alternative approaches, such as the ordinary equations method and the Q.R. factorization, will either fail or yield only one solution out of infinite possible answers. Although it is preferable to prevent over-parameterization, it may be challenging to identify in this specific case due to the high-dimensional nature of the data. We anticipate that the issues arising from our research will be balanced. The use of the pseudoinverse is a preventive measure. The Matlab function $pinv$ computes the Moore-Penrose pseudoinverse. In Matlab notation, the appropriate solution $\hat{\mathbf{x}}$ is obtained as follows:

$$\hat{\mathbf{x}} = pinv(\mathbf{A}) * \mathbf{b} \qquad (3)$$

```
{
    ‘type’: ’business’,
    ‘business_id’: (encrypted id),
    ‘name’: (business name),
    ‘neighborhoods’: [(hood names)],
    ‘full_address’: (localized address),
    ‘city’: (city),
    ‘state’: (state),
    ‘latitude’: latitude,
    ‘longitude’: longitude,
    ‘stars’: (rounded to 0.5-stars),
    ‘review_count’: review count,
    ‘categories’: [(local category)]
    ‘open’: True / False,
    ‘hours’: {
        (day_of_week): {
            ‘open’: (HH:MM),
            ‘close’: (HH:MM),
        },
        ...

    },
    ‘attributes’: {
        (attrib_name): (attrib_value),
        ...
        },
```

Fig 1. Yelp Dataset: JSON File Format for Business Data

When training our model to predict restaurant success, we employ cosine similarity to eliminate eateries that are not similar. Think of the characteristics of a restaurant as a vector of values that range from 0 to 1. Each vector element reflects the value of a specific feature related to that restaurant. Remember the dot product cosine identity:

$$\cos(\theta) = \frac{\mathbf{x}^T \boldsymbol{y}}{\|\mathbf{x}\|_2 \|\boldsymbol{y}\|_2} \tag{4}$$

or equivalently

$$\theta = \arccos \frac{\mathbf{x}^T \boldsymbol{y}}{\|\mathbf{x}\|_2 \|\boldsymbol{y}\|_2} \tag{5}$$

Where θ is the angle between the vectors **x** and **y**. It is a standard data-analysis practice to measure the `similarity' of two vectors based on the size of the angle between them (θ) [5].

1. Data Processing

We begin by filtering the complete JSON Yelp dataset, removing all Las Vegas businesses that are not restaurants. An additional data filter is applied to each model based on the restaurant type under consideration. When simulating Mexican cuisine, we apply an additional filter by restaurants that contain the word 'Mexican’ in their ‘categories’ field. Following this, the information is entered into a CSV file in which the columns contain numeric values that represent the attributes of the restaurants, and the rows correspond to the remaining restaurants in the data set. The attributes of continuous variables, including ‘price range’ and 'noise level,' are normalized using a scale ranging from 0 to 1. Discrete variables, such as 'location' or 'Wi-Fi available,' assign individual boolean fields to each option; a value of '0' indicates a 'no' or 'false' about that option, and a value of '1’ signifies a 'yes’ or 'true’ concerning that option. Observe that in the dataset, numerous restaurants failed to specify attributes. When this was logical, null fields were populated with an average or the default value to facilitate this analysis. If an excessive number of restaurants failed to answer for a given field, that field was omitted from the dataset due to data scarcity [25]. The corresponding CSV file is imported directly into our least square model as matrix $\mathbf{A}$. The average Yelp ratings for each restaurant are contained in the last column of this CSV file, which is imported into vector b for the least square modeling.

2. Identifying Significant Attributes

We conducted logistic regression on each model, utilizing the current restaurant category. We eliminate from the CSV file all columns that relate to attributes that are not significant. We select a p-value criterion that includes 12 to 18 columns in the CSV file [22]. The optimal range of remaining columns, rather than a single p-value, was the most effective for generating the best models. A p-value of $p > 0.11$ was selected for Chinese restaurants, resulting in 13 columns remaining in the CSV file.

3. Modeling

Four different methods were used for modeling and will be described below.

3.1. Average (Naive) Approach

The first (naive) approach was to simply use the unweighted mean of all the Yelp scores in a category as the prediction for future Yelp scores in that category. For example, the average Yelp score of all existing Mexican restaurants would be used to predict the Yelp score of a new Mexican Restaurant.

3.2 Least Squares (LS) Approach

The second approach, as recommended in [3], was to use Least Squares regression over all restaurants of a given category, using their attributes as dependent variables, as described in least square Modeling [27].

3.3 Cosine Similarity and Average

The third strategy relied on the concept of Cosine similarity, which was previously explained in Section II.3. This method employs cosine similarity to filter out restaurants not similar to the restaurant for which the Yelp score is being forecasted. The Yelp scores of the remaining restaurants are calculated by taking the average, as described in section II.3.1.

3.4 Cosine Similarity and Least Square (LS)

Similar to the previous method, this approach initially utilizes cosine similarity to narrow down the selection of restaurants. On this occasion, instead of using the average of the remaining Yelp scores as our forecast, we construct a LS model similarly as described in section II.3.2 [28].

4. Testing

To evaluate the effectiveness of our models and verify our hypothesis, we conduct elimination testing on each data point in our models. Put simply, we remove each restaurant from our existing data set (as it stands after section II.2) one by one, create a model using the smaller data set, make a prediction for the Yelp score of the restaurant that was deleted, and then compare our prediction to the actual data. The discrepancy between the anticipated and actual Yelp rating is recorded in a vector of residuals, which ultimately has a length equivalent to the number of restaurants accessible for that model [23]. The following section will investigate these residuals.

III. RESULT AND DISCUSSION

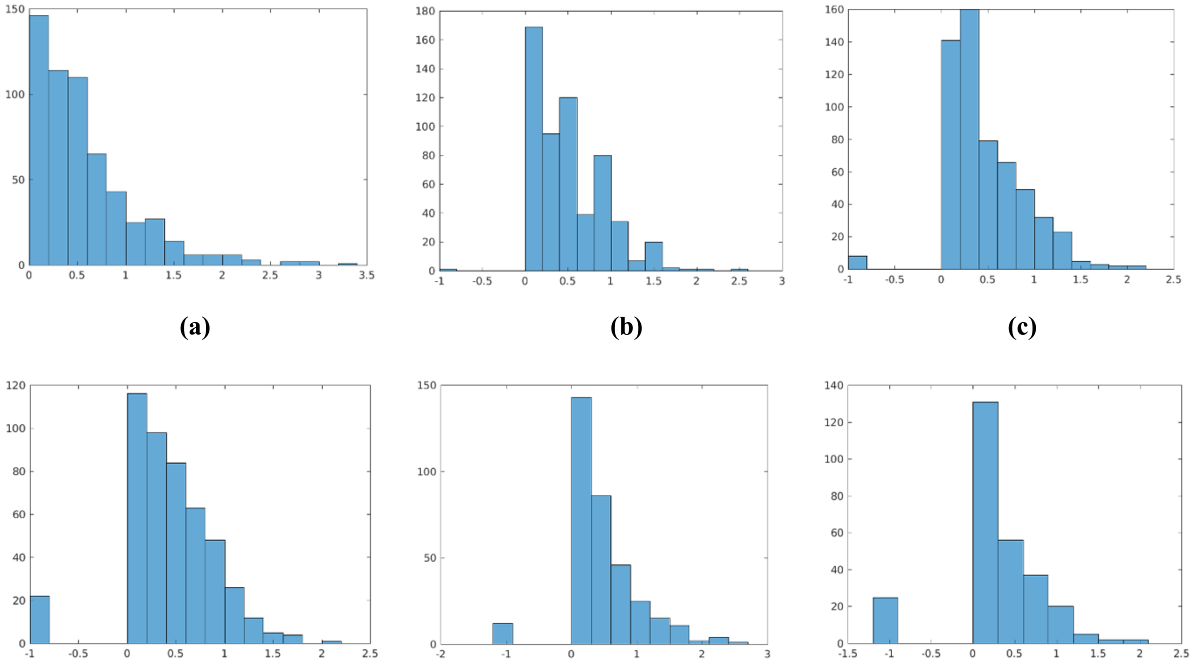
The models yielded the subsequent outcomes for Mexican, Traditional American, Chinese, and Italian restaurants, respectively, all located in Las Vegas. In order to calculate the residuals, a leave-one-out testing method was applied to each of the four models, as explained in Section II.4 and Section II.3.

Table 2 Displays the Average Residual, the identical data and histograms for Traditional American, Chinese, and Italian restaurants for each of the four models discussed in Section II.3. The histograms of the residuals are shown in Figures 2a, 2b and 2c. Note that -1 residual values indicate that after cosine similarity filtering, not enough data points remained to perform the analysis. These points were omitted from the residual vector when computing numeric results. It is essential to observe that the combination of cosine similarity filtering and LS modeling regularly achieves better results than any other method. The average difference between predicted and actual values is 0.4945, the distance between points in Euclidean space is 14.7910, and the mean squared error is 0.3893. This pattern persists across traditional American, Chinese, and Italian restaurants. The same data and histograms for Traditional American, Chinese, and Italian restaurants is presented in Figures 2d, 2e and 2f for the cosine similarity and LS method only. Note that for the American restaurants a p-value threshold of p < 0:2 was used, for the Chinese p < 0:11, and for the Italian p < 0:25.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23476

Table 2. Displays The Average Residual, the identical data and histograms for Traditional American, Chinese, and Italian restaurants

Model	Average	Least Square	Cos Sim, Avg	Cos Sim, LS	American	Chinese	Italian
Avg Res	0.5336	0.5745	0.5119	0.4945	0.5953	0.5153	0.4240
Eucl Norm	16.4298	18.2972	15.5237	14.7910	15.0144	12.8397	8.7855
MSE	0.4736	0.5873	0.4235	0.3893	0.5693	0.4951	0.3051



(a) (b) (c)

(d) (e) (f)

Fig. 2. (a) Histogram Of Residuals Using Average for Mexican Restaurants **(b)** Histogram of Residuals Using Average for Mexican Restaurants **(c)** Histogram Of Residuals Using Cosine Similarity And Average For Mexican Restaurants**. (d)** Histogram Of Residuals Using Cosine and LS For American Restaurants **(e)** Histogram of Residuals Using Cosine and LS For Chinese Restaurants **(f)** Histogram Of Residuals Using Cosine And LS For Italian Restaurants

IV. CONCLUSION

The findings of this research are that the application of cosine similarity followed using a least squares model consistently achieves better outcomes than similar methods. From these findings, we can deduce that restaurants, like one other, are indeed more reliable forecasters. Our four modeling efforts perform better than most miniature squares models across all restaurant categories. Creating distinct models for various sorts of restaurants might significantly enhance the accuracy of predictions. This result is supported by the finding that the statistically significant (with a low p-value) characteristics of restaurants vary among different restaurant categories. An illustrative instance is the variable 'Water service,' which holds statistical significance for Italian restaurants with a p-value of p = 0.031. However, the variable is not statistically significant for American restaurants, as indicated by a p-value of p = 0.282. Regarding location, the statistical analysis shows that the `Location_Westside' has a much higher p-value for American restaurants, indicating a stronger association. On the other hand, the `Location_summerlin' has a more significant p-value for Italian restaurants, suggesting a stronger correlation in that area. Our methods are considerably more practical for real-world use than earlier efforts in predicting Yelp scores. Nevertheless, their level of accuracy is still insufficient for practical applications. Nevertheless, our findings can offer valuable insights into the specific attributes that clients prioritize. This information may prove valuable to prospective restaurant proprietors and forthcoming modeling endeavors.

Our recommendation in the future research that the models in this field incorporate the following weighted regression uses cosine similarity as the weight instead of a binary inclusion or exclusion approach. Attempting to utilize non-linear models, earlier studies suggest that Least Squares (LS) may be the most optimal approach. Exclude establishments with a low number of Yelp reviews and chain restaurants from the model, as customer satisfaction does not correlate significantly with Yelp ratings. The Last, Utilizing the models to analyze different sorts of restaurants, various cities, and even other businesses inside the Yelp dataset. The examined literature illustrates the adaptability of linear regression methods in analysing and forecasting diverse elements of the Yelp dataset, encompassing review ratings, business attributes, and their interrelations with other factors. The meticulous implementation of regression analysis has yielded significant insights and enhanced conventional data sources in the realm of Yelp research. Moreover, the extensive applications of linear regression across several fields underscore its efficacy as a robust instrument for predictive modelling and data analysis.

Author Contributions: *Andrianshah Priyadi*: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision: Formulating ideas & research goals, writing the initial draft, reviewing & revising article, and supervising research activity. *Nelly Malik Lande*: Writing - Original Draft, Writing - Review & Editing: Formulating ideas & research goals, collecting data, writing the initial draft, reviewing & revising article. *Anita Faradilla*: Methodology, Investigation, Writing - Review & Editing: Developing and designing methodology. *Ma'arif Hasan*: Methodology, Investigation, Writing - Review & Editing: Developing and designing methodology. *Evi Widianti*: Writing - Review & Editing, collecting data, and reviewing & revising article.

All authors have read and agreed to the published version of the manuscript.

Funding: This project received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The **link** that support the findings of this project.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Andrianshah Priyadi: https://orcid.org/0009-0005-8037-1913
Nelly Malik Lande: https://orcid.org/0009-0000-8094-8827
Anita Faradilla: https://orcid.org/0009-0007-9288-7436
Ma'arif Hasan: https://orcid.org/0009-0000-9116-5043
Evi Widianti: https://orcid.org/0009-0001-2103-1186

REFERENCES

[1] Yelp dataset challenge. Available: [Online]. Available: http://dx.doi.org/10.4225/13/511C71F86-12C3

[2] S. National Restaurant Association et al., "Restaurant industry forecast," 2016.

[3] W. Farhan, "Predicting Yelp Restaurant Reviews," UC San Diego, La Jolla, 2014.

[4] S. W. I. Wahyudi, A. Affandi, and M. Hariadi, "Recommender engine using cosinesimilarity based on alternating least square-weight regularization," in *2017 15th Quantum Information Research*, 2017, pp. 1–6, **doi:** 10.1109/QIR.2017.8168492.

[5] Y.-H. Hu, K. Chen, and P.-J. Lee, "The effect of user-controllable filters on the prediction of online hotel reviews," *Inf. Manag.*, vol. 54, no. 6, pp. 728–744, 2017, **doi:** 10.1016/j.im.2016.12.009.

[6] A. Kong, V. Nguyen, and C. Xu, "Predicting International Restaurant Success with Yelp," Stanford University, 2016.

[7] M. Luca, "Reviews, reputation, and revenue: The case of yelp.com," *Harvard Business School NOM Unit Working Paper*, no. 12-016, 2011.

[8] L. Kwok and K. L. Xie, "Factors contributing to the helpfulness of online hotel reviews," *Int. J. Contemp. Hosp. Manag.*, vol. 28, no. 10, pp. 2156–2177, 2016, **doi:** 10.1108/IJCHM-03-2015-0107.

[9] E. S. Alamoudi and S. Al Azwari, "Exploratory Data Analysis and Data Mining on Yelp Restaurant Review," in *2021 National Conference on Computing and Communication (NCCC)*, 2021, **doi:** 10.1109/NCCC49330.2021.9428850.

[10] H. Li, C. (R.) Wang, F. Meng, and Z. Zhang, "Making restaurant reviews useful and/or enjoyable? The impacts of temporal, explanatory, and sensory cues," *Int. J. Hosp. Manag.*, vol. 83, pp. 257–265, 2019, **doi:** 10.1016/j.ijhm.2018.11.002.

[11] Y. Chen and F. Xia, "Restaurants' Rating Prediction Using Yelp Dataset," in *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, Dalian, China, 2020.

[12] M. R. D. Ching and R. de Dios Bulos, "Improving Restaurants' Business Performance Using Yelp Data Sets through Sentiment Analysis," in *2019 3rd International Conference on E-Commerce, E-Business and E-Government (ICEEG)*, 2019, pp. 62–67, **doi:** 10.1145/3340017.3340018.

[13] D. Keller and M. Kostromitina, "Characterizing non-chain restaurants' Yelp star-ratings: Generalizable findings from a representative sample of Yelp reviews," *Int. J. Hosp. Manag.*, vol. 86, 2020, **doi:** 10.1016/j.ijhm.2019.102440.

[14] J. Richards, S. Dabhi, F. Poursardar, and S. Jayarathna, "Poster: Leveraging Data Analysis and Machine Learning to Authenticate Yelp Reviews through User Metadata Patterns," in *2023 ACM Conference on Computer and Communications Security (CCS)*, 2023, **doi:** 10.1145/3565287.3617983.

[15] H. Nakahara, "A naive approach to program extraction," *Publ. Res. Inst. Math. Sci.*, vol. 25, no. 3, 1990, **doi:** 10.2977/PRIMS/1195173352.

[16] F. Pérez-González and C. Troncoso, "Understanding Statistical Disclosure: A Least Squares approach," in *Lecture Notes in Computer Science*, vol. 7384, 2012.

[17] R. Giri, ., Rymmai, and J. S. Saleema, "Book Recommendation using Cosine Similarity," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 3, 2017, **doi:** 10.26483/IJARCS.V8I3.2995.

[18] M. Sadikin and A. Fauzan, "Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset," *Eur. J. Electr. Comput. Eng.*, vol. 7, no. 6, 2023, **doi:** 10.24018/ejece.2023.7.6.583.

[19] Q. Xuan et al., "Modern Food Foraging Patterns: Geography and Cuisine Choices of Restaurant Patrons on Yelp," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 2, 2018, **doi:** 10.1109/TCSS.2018.2819659.

[20] M. Nakayama and Y. Wan, "The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews," *Inf. Manag.*, 2019, **doi:** 10.1016/J.IM.2018.09.004.

[21] C. Fu et al., "Link Weight Prediction Using Supervised Learning Methods and Its Application to Yelp Layered Network," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, 2018, **doi:** 10.1109/TKDE.2018.2801854.

[22] S. B. Hegde, S. Satyappanavar, and S. Setty, "Restaurant setup business analysis using yelp dataset," in *2017 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, **doi:** 10.1109/ICACCI.2017.8126196.

[23] T. Doan and J. Kalita, "Sentiment Analysis of Restaurant Reviews on Yelp with Incremental Learning," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, **doi:** 10.1109/ICMLA.2016.0123.

[24] R. Shen, J. Shen, Y. Li, and H. Wang, "Predicting usefulness of Yelp reviews with localized linear regression models," in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2016, pp. 189–192, **doi:** 10.1109/ICSESS.2016.7883046.

[25] S. S. Kumar et al., "Unveiling Patterns and Enhancing Recommendations: A Novel Regression Analysis Approach for Yelp Dataset," in *2023 International Conference on Next Generation Electronics (NEleX)*, 2023, **doi:** 10.1109/NEleX59773.2023.10421042.

[26] N. Asghar, "Yelp dataset challenge: review rating prediction," *arXiv preprint*, 2016, **doi:** 10.48550/arxiv.1605.05362.

[27] E. Anenberg, C. Kuang, and E. Kung, "Social learning and local consumption amenities: evidence from yelp*," *J. Ind. Econ.*, vol. 70, no. 2, pp. 294–322, 2022, **doi:** 10.1111/joie.12291.

[28] M. Dolatabadi et al., “Cognitive sequential dependencies in the wild: sentiment analysis approach,” *arXiv preprint*, 2020, **doi:** 10.31234/osf.io/4mw8c.

[29] Y. Shen et al., “Using social media to assess the consumer nutrition environment,” *Public Health Nutr.*, vol. 22, no. 2, pp. 257–264, 2018, **doi:** 10.1017/s1368980018002872.

[30] M. Rahman, B. Carbunar, J. Ballesteros, and D. Chau, “To catch a fake: curbing deceptive yelp ratings and venues,” *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 8, no. 3, pp. 147–161, 2015, **doi:** 10.1002/sam.11264.