Performance Improvement of Restaurants using Yelp Dataset

Yashas Kadambi
Department of Computer Science
PES University
Bengaluru, India
yashasks@pesu.pes.edu

T Vijay Prashant
Department of Computer Science
PES University
Bengaluru, India
tvijayprashant@gmail.com

Manav M
Department of Computer Science
PES University
Bengaluru, India
manav.gowda.0@gmail.com

Vaibhav Jamadagni
Department of Computer Science
PES University
Bengaluru, India
vaibhav.jamadagni@gmail.com

***Abstract*—Restaurants have always played an role in the business, social, intellectual and artistic life of a thriving society. In the modern world this results in restaurant being an important part of economy of many countries with contributions to both to the GDP and to the employment opportunities. There is stiff competition in the restaurant industry and often a lot of restaurants struggle to keep business and close down. This paper proposes to detect the restaurants which are going to close down in the near future and also provide ways to improve their business based on the features of successful restaurants. It is implemented using Yelp dataset, the checkins are used to classify the restaurants into open or closed, the reviews are used to find amenities, features, cuisines etc that a restaurant can offer to improve its business. Improving quality of restaurants increases the quality of life of the general population along with increase in GDP and employment opportunities.**

I. Introduction

Restaurant industries is large factor in determining a countries Gross Domestic Product GDP [1]. In the United States restaurant industry contributed about 4% to GDP. In India the restaurant industry is dominated by the unorganised sectors like food carts and street vendors. How ever in recent times there is significant shift to organised sector like fine and casual dining with casual dining having higher growth. This shift to organised sector is expected to grow as the preferences of the consumer and increase in middle class population. In addition these type of restaurants provide employment to people having higher educational qualification than the unorganised counter parts, which in turn contributes to the employment of middle class population as qualities like spoken English is often found in the middle class workforce quite easily. Restaurants also play an important role in employment as around 10% of the total employment is generated by the restaurant industry in the US [2].

Consumer reviews have become ingrained in our decision-making process. Before making a purchase, most people look at online reviews. When it comes to picking a set of services or products, consumers in the world today have a wide range of options further enhanced by access to a vast database of reviews, ratings, and general information provided by the community. Yelp is an online review site which has risen to popularity which is leading to an increase in data of peoples preferences and personalities. This dataset can also be leveraged to help business owners to improve their businesses [3]

Restaurant owners want to know whether they will succeed in the future. Yelp's rating is one of the most important indicators. According to Dimensional Research [4], 90 percent of customers say they make purchases based on online reviews. Having a positive review indicates a restaurant offers formidable customer service, a fun environment and high-quality food. Another study [5] that positive reviews affect consumers' trust in a local business, with 72 percent saying positive reviews help in entrusting local businesses. while poor reviews can drive away prospective customers,it also can be used to improve customer service and experience.

Yelp ratings have a profound effect on the success of businesses as "an extra half-star rating causes restaurants to sell out 19 percentage points more frequently" (an increase from 30% to 49%) [6]. To make online reviews more useful, Yelp alone contains more than 70 million reviews of restaurants, barbers, mechanics, and other services, and has a market capitalization of roughly four billion dollars [7]. Not only The dataset undergoes yearly updation it also flags suspicious and fake reviews using filtering algorithms. These reviews if put use can improve the restaurant industry to better cater to consumers, thus helping the economy and employment opportunities.

Successful restaurants have the luxury to make use of part of their income to research current trends and enforcing changes in their restaurants to better cater to customer needs. Struggling restaurants on the other hand need to improve their restaurants without expensive methods. By making use of public datasets like the Yelp dataset and finding out the difference between the struggling restaurants and the popular restaurants is a way to help the struggling restaurants do better. Ultimately

the user benefits as more and more restaurants become popular and hence increasing competition thus driving the overall prices down and enhancing the customer experience.

The goal is to help struggling restaurants using the Yelp dataset. The proposed model implements the following to achieve this. By analysing the footfall of closed restaurants predict whether the struggling business will go out of business, using which these businesses can take decisive actions using the these parameters.Based on the locality identify the popular restaurants and suggest cuisines and amenities which are important for the clientele. Using sentiment analysis the reviews are analysed to improve the quality of restaurants. .

All the existing models lack in the fact that they completely disregard the locality and the culture of the clientele of restaurants. The proposed approach is expected to perform better as restaurants of different neighbourhoods within a city have very different need of amenities as the average income of people living in different localities differ significantly. For example, restaurants in suburbs offering parking space do well, while restaurants in a busy city centre do well without offering any parking at all. The locality has a significant impact on the cuisine and amenities offered by the restaurants. For example, Cafes doing well in the city centre offer WiFi, while a cafe in a locality of senior citizens can do well without offering WiFi.

In this paper we propose to analysing the reviews of restaurants of a particular locality. the locality is generalized to cities as a reasonable estimate. The section II goes in detail about all the work which has been done on the yelp dataset, section III explains the the proposed solution, section **??** explains the inferences made form the model.

II. Literature survey

the paper [6] explain the latent subtopics discovered from Yelp restaurant reviews using a Latent Dirichlet Allocation (LDA) algorithm. The aim is to point out the demand of customers from a large number of reviews, with high dimensionality. For restaurants, these topics provide useful insight into what customers are looking for. The paper presents the breakdown of hidden topics throughout all reviews, predicts stars based on hidden topics discovered, and extends findings to temporal information about restaurants peak hours.Overall, several insights are found which are useful to restaurants. Using the LDA algorithm where various signals from the user reviews and the spatial data are combined and the demand of customers from a large number of reviews are analyzed in a more dimensional manner and provide meaningful insights to restaurants about what their customers care about in order to increase their Yelp ratings.

The papers [8]–[10] propose multiple models for restaurants' rating prediction using Yelp dataset. The paper [8] predicts rating in 2019 based on the collected information from 2018. Both non-text features and text features are applied to provide suggestions to restaurants on yelp, aiming to work on a user-based analysis. Regression model, Naive Bayes, Decision Tree, and Neural Network models were used and a best accuracy is 82.50% was observed. The paper [9] predicts a business' rating based on user-generated reviews' texts alone. This not only provides an overview of plentiful long review texts but also cancels out subjectivity. Selecting the restaurant category from Yelp Dataset, a combination of three feature generation methods as well as four machine learning models is used to find the best prediction result. The approach is to create bag of words from the top frequent words in all raw text reviews, or top frequent words/adjectives from results of Part-of-Speech analysis. Our results show Root Mean Square Error (RMSE) of 0.6. The paper [10] predicts the rating for a restaurant from previous information, such as the review text, the user's review histories, as well as the restaurant's statistic. Three machine learning algorithms are used, linear regression, random forest tree and latent factor model, combining with the sentiment analysis. The best model for predicting the ratings from reviews is found to be the random forest tree algorithm. The above papers predict the rating of the restaurant based on the user reviews and ratings alone.The results can be improved upon by considering local factors like food preference varies a lot from region to region.

Singhu Hegde et, al. [11] provides the issues associated with setting-up of a new restaurant business are addressed. To strategize a new restaurant, a restaurant business framework which comprises of 3 most important tasks namely high frequency attributes, most crowded day and location of the restaurant are required. First, the features/attributes of the restaurants in which the customers are most interested in are discovered and those facilities and services should be provided to increase profits. Also finding the days of the week when the restaurants are heavily crowded helps in the best recipes and be made available on those days. Finally, since location has a profound effect on the success of a restaurant business, location to be the most important to know the nearby restaurants and their facilities before coming up with a new restaurant business. We improve upon the findings of this paper by helping strategize business models not only for newly opening restaurants but also for struggling ones.

III. Methodology

A. Data

This project makes use of the all the datasets from the Yelp Dataset Challenge. The datasets included in this challenge are business, review, checkin, tip and user in the form of separate json objects. The business dataset contains attributes such as the business name, the unique identifier, the name, the address, the city, the stars, the count of reviews, the category of the business, features offered by the restaurant and the open attribute indicating whether the business is currently open. A check-in dataset containing the business id and a list of time stamps for all customers checking in. A review dataset which contains the identifiers for review, business and user, stars given to reviews, a count of the attributes funny, useful, and cool, review text, and timestamp. For simplicity we mainly use these three json objects. Moreover, we only

examine businesses that are "restaurants" and only reviews that are associated with these restaurants.

B. Pre-Processing

The business dataset is processed and the features offered by the restaurant are separated into a separate csv file and the rest of the attributes are accumulated into a csv file. The checkin json object is processed into hourly check-in count per day per business and stored in another csv file. The latter two csv files are then merged and two new datasets are created with restaurants that are currently open and restaurants that are closed. These two datasets form the main basis for all the further analysis performed and hence will assist in deducing the necessary inferences.

An initial visualisation of the top cities with the most number of reviews is performed and based on this, the city 'Las Vegas' is chosen for demonstration purpose. To facilitate this all the dataset are filtered for Las Vegas wherever necessary.

In order to run the LDA model, the data from the two datasets are preprocessed with the Gensim CPython library, which converts sentences to words. The stopwords in these words are then removed using the nltk stopwords corpus before the corpus can be used to form bigrams and trigrams. SpaCy is a memory-efficient CPython library that excels at large-scale information extraction. This library is used for the lemmetization and stemming words from the corpus. Lemmatization and stemming of words from corpora are performed using this library. These corpora are then converted into respective dictionaries and the term document frequency is determined. This is then used as input into the LDA model.

C. Models

a) Checkin Analysis: The filtered Las Vegas checkin data of restaurants is taken and checkin for all the restaurants ($\sim$15000 restaurants) is aggregated over hours for each day of week is plotted as in 1. It can be seen that pattern over the weekdays remains approximately same also the no of checkins is just over 60000 which is around 4 checkins per day for each restaurant. This shows the sparsity in data. since the hourly pattern of the checkins is same for all weekdays, it is a reasonable approximation to aggregate it over weekdays. After all preprocessing we have checkins at a particular hour for all days of each restaurant. This data is used to fit different models including decision trees, multi layer perceptron etc, to classify whether the hotel with a particular amount of checkin is gonna succeed or not. How ever this alone is not nearly enough to prove that a restaurant is going to remain open or closed.

b) Latent Dirichlet Allocation(LDA): Yelp is a review platform that connects people with local businesses. This platform allows customers to find businesses and helps businesses understand their customers' needs to better serve them. This platform heavily relies on crowd sourced plain text reviews. The restaurant may not be able to figure out what it's customers desire from a large number of reviews, the challenge is identifying what users care about the most when they write reviews, and ultimately determine what restaurants are doing to receive good ratings. In order to gain new insights, it is beneficial to identify latent topics and subtopics in Yelp reviews.One can deduce a positive or negative feedback on each topic and subtopic. This can be achieved by performing various techniques to extract word features by keeping the initial context and sentence structure like bag of words, Term Frequency-Inverse Document Frequency(TF-IDF). On this corpora we can use various Machine Learning Techniques such as K-Means, LDA, LSA. LDA model is considered as it is faster on the cleaned corpus containing words with TF-IDF. Latent Dirichlet Allocation (LDA) is an unsupervised Bayesian generative model for text which is a topic modeling technique to extract topics from a given corpus. LDA assumes that a corpus of text documents cover a collection of K topics. This method helps us in extracting the hidden topics in the reviews of the restaurants. LDA generates summaries of topics in terms of the discrete probability distributions over words, and it further infers per-document distribution over topics. The cleaned and processed corpora in the previous section is used as inputs to the LDA Multi Core Model with α = 0.01 and β = 0.9 and number of topics = 8. These values gave the best result for the dataset provided. Although a full explanation of the LDA algorithm is beyond the scope of this paper, an intuition to where and how it is given in the result section.

IV. Experimental results

a) Checkin Analysis: The data was split into 80-20 train and test split and fitted with various models including MLP, linear regression, logistic regression, Adaboost Ensembele model. All of the above models achieved an accuracy of 80% this shows that about 20% data is getting misclassifed. We can make an argument that all the instances which are open and getting misclassified as closed are those restaurants which are struggling and are about to close down. Although this does not guarantee that those restaurants which have very less customers close down for example a restaurant which gets it majority of business elsewhere for example catering, hosting events, etc, it provides a reasonable basis for those restaurants which are predicted as struggling to improve. Helping a restaurant improve even though its not closing down does not in any way harm the restaurant. it only increases the competition to deliver the best to customers. Although predicting whether a restaurant will open or close can be improved with a timeseries data of all the checkins over a year or a month which is not recorded by Yelp.

b) Latent Dirichlet Allocation(LDA): Based on an initial wordcloud analysis of the user reviews of the open restaurants, it appears that the reviews tend to emphasize the service, food, friendliness of the staff, and the ambience of the restaurants.This can be seen in Fig.2

In the wordcloud for the closed restaurants, it appears that the reviews tend to focus on the food of the restaurants and rarely on the service they provide. A few negative reviews can also be seen in the Fig.3

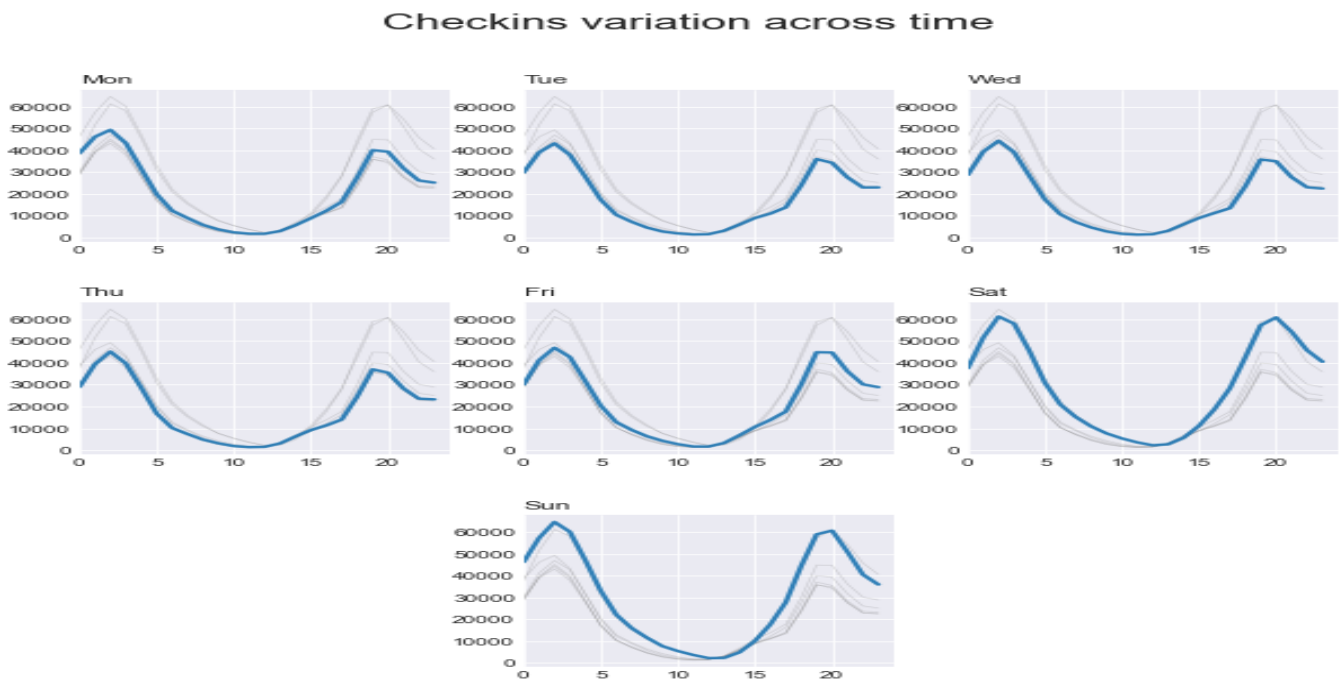

Fig. 1. Aggregated checkin time over hours for every weekday

Fig. 2. Word Cloud of the Open Restaurants

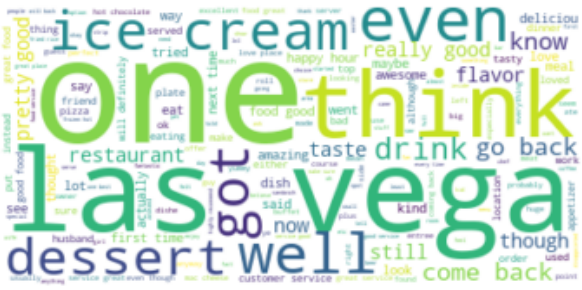

Fig. 3. Word Cloud of the Closed Restaurants

Possibly one of the reasons restaurants closed down is that even with great food and tasty desserts as the service was subpar for Las Vegas standards.Las Vegas in one of the major tourist attraction in the US an people usually go to Las Vegas to relax. There is a strong customer preference for good and fast service, as well as great ambience, according to the wordcloud.

The LDA achieved a coherence score of 0.4 which is considered good.The coherence score measures how much all the reviews support each other. It is not a measure of the model quality rather the data quality. pyLDAvis library is used to visualise the results of the LDA model. The plots in Fig.4 and Fig.5. The plot shows a flattened 2 dimensional figure of our data. The plot on the left shows the distance measure between the topics i.e how related is each topic. The graph on the right shows the frequency of occurrence of the term. the red chart shows the frequency of occurrence of the selected topic (the red circle on the left). The visualisation is highly interactive and is extremely hard to demonstrate as pictures. the lambda value the slide bar can be adjusted to view the most relevant words to a particular topic or the whole data in general. Higher lambda value shows frequency of occurrence of data in the topic selected. Lower values shows wrt to the entire topic.

The Fig. 4 is a visualisation of LDA of all the restaurants which are open in the city of Las Vegas. Adjectives like great, amazing, love friendly etc can be mapped to other terms in the chart for example great food, amazing food, friendly or amazing staff and service etc. We can be sure that the reviews are positive due to lack of negative words and the no of occurrences being in the region. this is further supported by the word cloud. Further all of these words are ordered in decreasing order of occurrence for 1 topic. similarly all the other topics can be examined to get additional information about the open restaurants in Las Vegas.

The Fig. 5 is a visualisation of LDA off all the restaurants

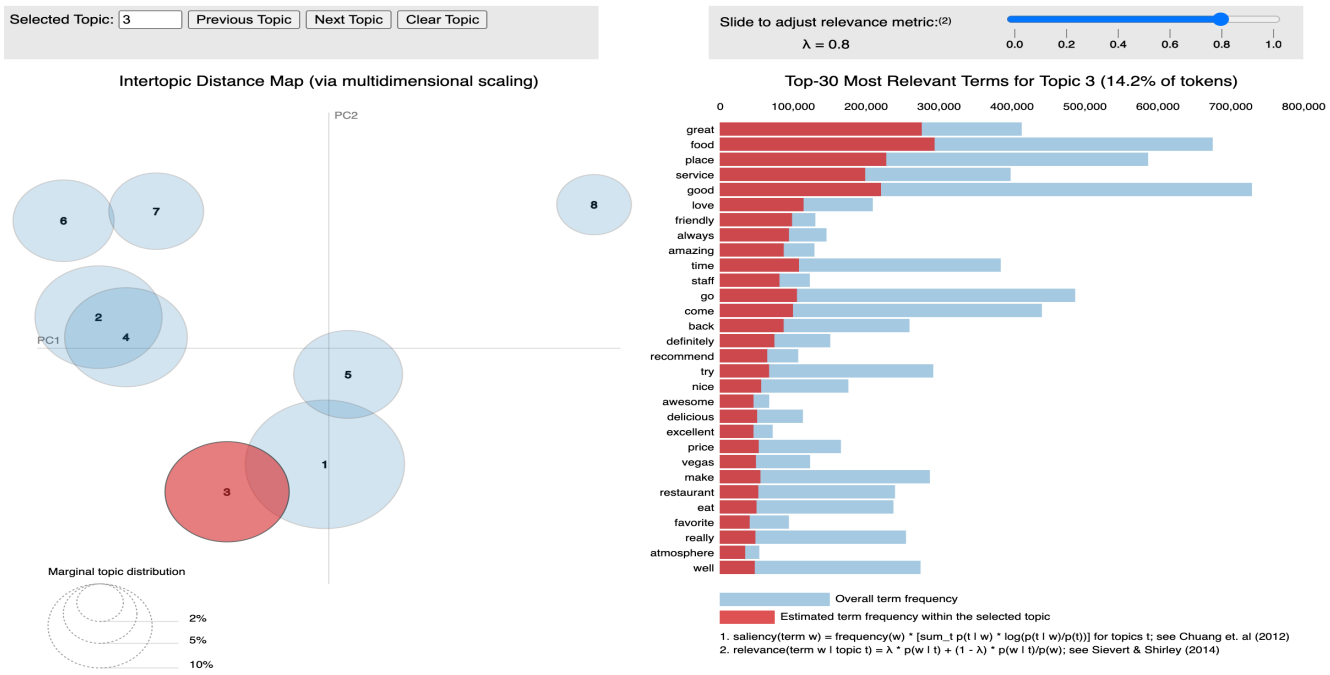

Fig. 4. LDA of the Open Restaurants

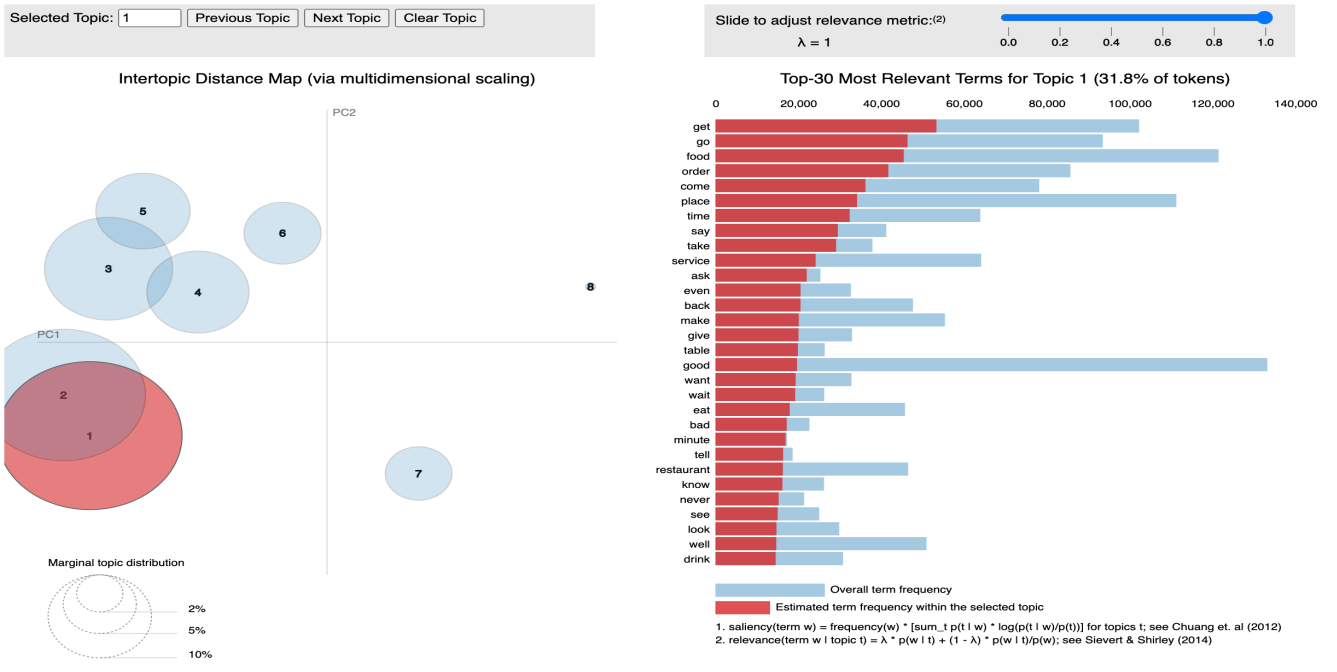

Fig. 5. LDA of the Closed Restaurants

which are closed. in this plot negative words like take time, bad, never, wait etc tells what these restaurant is lacking. These terms can be attributed to the reason for closure. Also the comparison of topics which lie in same region in left graph provides more assurance of the accuracy of the comparison between open and closed businesses as essentially the same data is fed to the model. After encountering some reasons for closure in one of the topics ,other topics in the vicinity should be analysed thoroughly. In the plots depicted the topics surrounding the selected topic confirms the bad service, long wait times and also shows that many restaurants did not offer takeout. By reducing lambda values additional factors like amenities and availability of drinks, ice cream, etc can also be seen. LDA hence provides concrete measures to find qualities which makes a restaurant succeed.

V. Conclusion

With the change in lifestyle of the world and working women becoming the norm many are choosing to order from restaurants frequently. This is resulting in huge growth of the restaurant industry. In countries like India this boom in restaurant is just in starting phases. It can be observed that ordering from restaurants on daily basis is already norm in western cities like New York especially when it comes to the current youth. Successful restaurants and restaurant chain have very little problems when it comes to adapting to new clientele. The revenue of these restaurants is good enough to invest in improving their restaurants. This will result in monopoly of few restaurants. Monopoly of anything usually decreases innovation. Hence to help struggling restaurants improve their business we proposed the above methods of detecting struggling restaurants and help them improve by comparing the reviews of restaurants which are closed and open restaurants to gain insight into what different things are the successful restaurants doing which struggling restaurants can adopt. These insights are more meaningful when they are based on a locality as requirements of clientele vary from locality to locality. The struggling restaurants are identified using the number of checkins and binary classification models. The feature prediction is done using sentiment analysis of reviews using LDA. We can use these two models to help struggling restaurants to improve their business and avoid going out of business.

The proposed solution has a lot of limitations when it comes to implementing in the real world. The prediction of struggling restaurants using yelp dataset is possible using very minimal data and hence high accuracy cannot be achieved. The lack of time-series data over long periods for a restaurant significantly affect the classification results. However the misprediction(false positive) only result in improving quality of restaurants which is always helpful to the clientele in the end. Using the locality as entire city reduces performance or the confidence in the identified features as some features which offered by restaurants of a particular neighbourhood are obsolete in other neighbourhood but is essential in that particular neighbourhood. Further analysis like this is quite easy as only the filter has to be adjusted filter only one neighbourhood rather than the entire city. The filtering done is a basic filtering based on fixing a random point in the city of Las Vegas and taking all restaurants in a 5 mile radius. Usually neighbourhood are rarely in perfect clusters and this reduces the performance as well. Significant analysis of data has to be done after the LDA with respect to a particular restaurant including comparing features, amenities, etc. Although the analysis will give the most probable reasons why the restaurant is not performing well, other factors like restaurant location have a huge impact which cannot be modeled in the classification problem. Even after identifying the problems it is up to the restaurant management to successfully find and implement a solution, the model merely offers the problems found in restaurants.

References

[1] U. ERS, "Food sectors and the economy," 2019.

[2] S. Singh, "What is happening with the unorganised restaurant sector amid pandemic," 2021.

[3] smcin, "Restaurant employee statistics," 2020.

[4] F. Abdullah, A. Abdurahman, and J. Hamali, "The dimensions of customer preference in the foodservice industry," *Verslas: teorija ir praktika*, vol. 14, pp. 64–73, 03 2013.

[5] B. Local, "Local consumer review survey," 2020.

[6] J. Huang, S. Rogers, and E. Joo, "Improving restaurants by extracting subtopics from yelp reviews," 2014.

[7] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

[8] 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)*, 2020, pp. 113–117.

[9] M. Fan and M. Khademi, "Predicting a business star in yelp from its reviews text alone," 01 2014.

[10] M. Yu, M. Xue, and W. Ouyang, "Star prediction for yelp dataset," 2015.

[11] S. Hegde, S. Satyappanavar, and S. Setty, "Restaurant setup business analysis using yelp dataset," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 2342–2348.