國立臺灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management
College of Management
National Taiwan University
Master Thesis

基於使用者生成內容進行餐廳選址排序

RESPECT: Restaurant Placement Ranking via Leveraging User-Generated Contents

江雨柔

Yu-Jo Chiang

指導教授：魏志平 博士

Advisor: Chih-Ping Wei, Ph.D.

中華民國 112 年 7 月

July 2023

致謝

時光荏苒，碩二的時間走的飛快，首先感謝在我茫然無措進入資管所時，將我納入羽下的魏志平教授。老師真的就像大家的大家長一樣，時不時關心以及帶領我們一起學習或是研究。耐心的教導、不厭其煩地叮嚀，以及 24 小時隨時解答論文研究的疑惑，真的是我進資管所最大的幸運。最重要是老師總是能從容不迫的解決很多問題，讓常常手足無措不知如何是好的我吃了不少定心丸，還有偶而也會開示我們人生還有很多選項不只是窮忙窮碌。

感謝我的同學們：海因、佑甄、佩琳、宇雋以及璟耀。兩年一起努力的感覺真好，碩一一起為了課業、實習在實驗室拼搏，碩二為了碩論埋頭苦幹，當然還有那些快樂的小確幸時光，讓研究所的路沒有那麼辛苦，現在回想都覺得奢侈又幸福，很高興能認識一群又棒又厲害的小夥伴們。特別感謝博班的學長們以及大學長：仁瀚、晉華與沛原學長，在我寫論文這段期間給予了莫大的幫助，不論是研究法的討論或是單純解惑。尤其仁瀚學長真的是盡心盡力，讓我的研究每次卡住的時候總是又能找到下一個出口，謝謝學長花如此多的時間和我一起討論一起分析，讓我的論文得以邏輯通暢且完整。

感謝我的家人們，無私的栽培讓我得以無憂無慮地完成學業。感謝我的小老妹，默默的支持與守護讓我求學時能身無旁騖的把我的這份學業完成。感謝我周遭的新朋友、老朋友們陪我、鼓勵與支持以及一起完成研究所學業，研究所固然辛苦，但也因為有你們而顯得值得。

最後，開始著手進行研究到最後完成這份論文，是我截止目前為止最浩大的工程，要感謝的人絕對遠大於上述，在此統一感謝在這一路上給我信心、見解、指導的同學、老師、朋友、同事與上司，讓我這一路走來真的不孤單。回顧至今，這些磨練與經歷都讓我更上一層樓；這些陪伴都讓我更加成長茁壯。謝謝大家一起見證，未來將會帶著這份祝福與鼓勵繼續努力。

江雨柔 謹識

于臺灣大學資訊管理學研究所

中華民國一百一十二年七月

摘要

在餐飲業日益普及的背景下，如何在百家爭鳴中脫穎而出成為重要的商業議題，策略性地選址規劃變得格外關鍵。除了從供給與需求數據作為決策考量外，在社群媒體發展快速的年代，分析使用者生成的內容 (UGC) 來了解消費者的在乎的面相以及對應的滿意度更是重要的資訊管道。透過屬性層級情感分析（ABSA）技術的不斷改進，我們現在能夠更快且更真實地去了解用戶的反饋與需求。

在預測新店面適合的位置時，我們使用了多種特徵，其中包括描述候選位置以及餐廳的表現和競爭力的相關信息。然而，傳統的競爭力計算方法卻常常依賴於商業資料、商店密度或評分等單一因素。為了克服這些限制，我們提出了一種全新的方法，採用了 ABSA 技術來分析客戶的評論和反饋。此外，我們的競爭者分析採用了社群檢測方法，使得競爭者群組得以更為準確地被識別。這種方法與傳統的基於類別的競爭者分析相比，能夠提供更精確和深入的競爭環境分析。

經過一系列實驗，我們的研究結果顯示所提出的方法的有效性，並進一步解釋所得出的分析結果。我們的方法不僅能夠顯著改進餐廳選址決策的準確性和相關性，還展現了利用 ABSA 技術來制定更細緻、更數據驅動策略的潛力。總結，本論文的貢獻在於提出了一種全新且高效的方法，將情感分析和競爭者分析有機地結合，並採用 ABSA 技術來實現對餐廳選址的精細化評估。我們的研究將為餐廳業務和選址規劃提供寶貴的洞察，幫助餐廳業主和相關利益相關者做出更明智的決策，以在競爭激烈的行業中脫穎而出。

關鍵字：店家選址、社群分析、使用者生成內容、屬性層級情感分析、深度學習

ABSTRACT

In the context of the increasingly popular catering industry, how to stand out in the intense competition has become essential, hence, selecting an optimal location for opening brick-and-mortar restaurant became even more indispensable. Besides relying on supply and demand data as decision-making factors, with social media rapidly developing, analyzing user-generated contents (UGC) to understand consumers' preferences and satisfaction has become a crucial information. Through Aspect-Based Sentiment Analysis (ABSA) techniques, we now have a faster and more authentic means of comprehending user feedback and demands.

In predicting the suitable location for new restaurant establishments, we employ multiple features, descriptions of candidate locations, restaurant performance, and competitiveness-related information. However, conventional methods of calculating competitiveness often rely on single factors such as commercial data, shop density, or ratings. To overcome these limitations, we propose a novel approach that analyzes customer reviews and feedback. And also utilize review with community detection to offer more precise and in-depth analysis of the competitor landscape.

The research results demonstrate the effectiveness of the proposed method and provide explanations for the derived outcomes. The proposed approach not only significantly enhances the accuracy restaurant placement decisions but also showcases the potential of using ABSA techniques to formulate more refined and data-driven strategies. In conclusion, the study offers valuable insights for restaurant businesses and site planning, aiding restaurant owners and relevant stakeholders in making more informed decisions to excel in the fiercely competitive industry.

Keywords: Store site selection, Community detection, User generated content, Aspect-based sentiment analysis, Deep learning

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

Location of a store is a crucial factor in determining its success and profitability when opening a new branch. As the age-old maxim goes “No matter how good it is offering, merchandising, or customer service, every retail company still has to contend with three critical elements of success: “location, location, and location”(Taneja, 1999). Surveys conducted by the Small Business Administration (SBA) and other organizations indicate that poor location is one of the primary causes of business failure in America[1]. Fox et al. (2007), Li and Liu (2012) all stated that even though there are many factors influencing the success or failure of a store, one factor still manages to stand out, which is location. A survey conducted by Foursquare[2] stated that when the store location is not accurately determined, 46% of the businesses gained higher customer acquisition costs, 40% lost market shares and 42% had lower quality of customer experience.

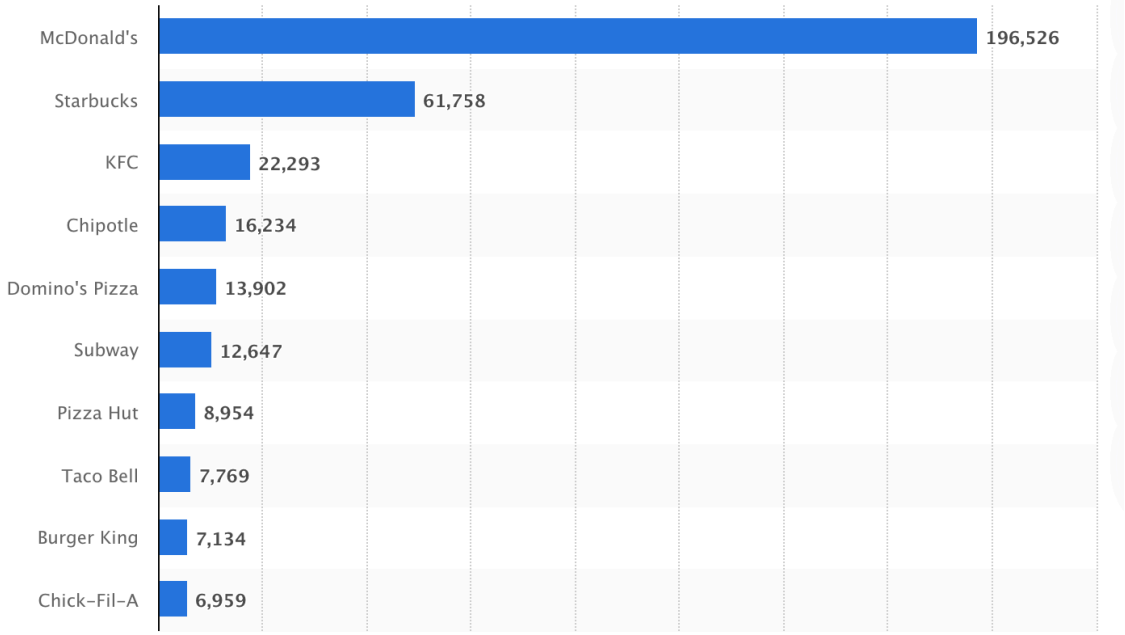
It is acknowledged that the level of success of a business is heavily dependent on operating the business from the right location. The key determinant is demand, hence, defining an attractive site location that would generate high demand and would minimize costs is the ultimate goal. Shop owners or managers would want to choose a location where people have the inclination to shop, the time to spend in your store, and the means to buy your product. Features can include the state of the environment, the number and location of competitors as well as the degree of their success, the cost of renting the premises, and a wide range of indicators related to future attendance of the establishment

[1] Dirkes, D. (2014) Business Failure Stats & Reasons Summary. Linkedin (Sep 21, 2014) Available at https://www.linkedin.com/pulse/20140921225654-8175758-business-failure-stats-reasons-summary (Retrieved on July 23, 2023).

[2] Forrester (2022) Location Intelligence Builds Innovation, Optimization, And Better Customer Experiences. (July, 2022) Available at https://location.foursquare.com/wp-content/uploads/sites/2/2022/08/Forrester_Opportunity_Snapshot_-_Report.pdf (Retrieved on July 23, 2023).

(Hernandez & Bennison, 2000). In conclusion, store positioning is a complex and crucial process that requires comprehensive consideration of various factors. Through scientific analysis and strategic planning, we will be able to identify the ideal location best suited for store development.

In the catering industry, by the end of 2022, McDonald's had around 40,275 locations worldwide[3], Starbucks with 35,711 stores[4] and KFC with over 27,700 locations[5]. With so many stores to choose from, this is not an easy task to do. Looking at Figure 1-1we can see that they still manage to stay on top 3.



Source: Statista (2023)

Figure 1-1 10 most valuable quick service restaurant brands worldwide in 2022[6]

McDonald's has an extensive and detailed decision system for how to locate the store which includes demographic databases and population statistics-based marketing

[3] McCain A. (2023) 22 Mcdonald's Statistics[2023]:Restaurants Counts, Facts, and Trends. Zappia (March 21, 2023) Available at https://www.zippia.com/advice/mcdonalds-statistics/ (Retrieved on July 23, 2023).

[4] Statista Research Department (2023) Number of international and U.S.-based Starbucks stores from 2005 to 2022. Statista (Nov 28, 2022) Available at .https://www.statista.com/statistics/218366/number-of-international-and-us-starbucks-stores/ (Retrieved on July 23, 2023).

[5] Statista Research Department (2023) Number of Kentucky Fried Chicken restaurants worldwide 2010-2022 Statista (Mar 31, 2023) Available at https://www.statista.com/statistics/256793/kfc-restaurants-worldwide-by-geographic-region/ (Retrieved on July 23, 2023).

[6] Statista Research Department (2023) Brand value of the 10 most valuable quick service restaurant brands worldwide in 2022 Statista (May 9, 2023) Available at https://www.statista.com/statistics/273057/value-of-the-most-valuable-fast-food-brands-worldwide/ (Retrieved on July 23, 2023).

research to target consumers and analyze competition within a radius of five to seven miles and find out if the location is suitable for franchises. KFC also sets up an estimating method to score the locations or business districts by its population, consumption power, revenue of the business district and the convenience of transportation. After evaluating through all these metrics, KFC chooses the highest-scored places and distributes branches in buildings, supermarkets, bookstores, trade centers, etc. Atlas provides Starbucks with data on consumer demographics, population density, income levels, auto traffic patterns, public transport stops, and the types of stores/businesses in the location under evaluation. Thus, Starbucks is able to project foot traffic and average customer spending in a given location, therefore helping Starbucks to determine the economic viability of opening a store in that spot. In general, past literature and business nowadays confirm a statistically significant relationship between site location and business performance, and continuing efforts have been made to determine emerging factors affecting this relation.

The problem of new store location selection has been well studied. When opening a new store or branch, many factors should be considered. By investigating through the existing literature, we can separate these methods into two parts: **survey-based approach** and **data-analytics-based approach**. Survey-based methods are those that make use of surveys and expert experiences, which are too time-consuming and need to be more scalable in this fast-changing world. On the other hand, data-analytics-based methods collect data from all sources, retrieving data with both volume and velocity. These methods can be divided into several features by the data source, namely: geo-economics-based, location-based, and social-media-based features. Geo-economics-based features take advantage of geographic, demographic, and transportation perspectives to describe places or restaurants from consumers' perspectives rather than just figures generated by

the store or restaurant itself. A clearer picture of consumer behavior at a specific location, allows us have a better understanding of consumers and segmentation, leading to a more accurate prediction. Location-based features include data from application, such as Foursquare or maps' point of interests (POI) check-ins depicting the factors with customer flow showing the popularity of one location with spatial and temporal information. Most of the works tackle the problem with this approach (Fox et al., 2007; Li & Liu, 2012). Social-media-based features require data from social media applications or review websites. By leveraging the user-generated contents, we can now take users' opinions into consideration. We can also gain insight into social validation such as likes and comments, these validations not only signal customers but also affect their behavior. These features typically focus on capturing information related to a user's location, activities, interactions, and preferences, scaling the restaurant from the consumer perspective, so that we can have a finer view of the performance of the restaurant.

1.2 Research Motivation and Objectives

According to the features of the existing work used in prior studies that consider reviews as an input, view review counts as a way to measure popularity or success, instead of using the aspects or sentiments inside them. Wang et al. (2016) is the first to leverage aspects and sentiment of the review. The aspect gained from LDA are in document level, meaning the topics are presented by the words in the documents, resulting in too fine grained or diverse aspect, making it hard to compare the performance between them, hence a higher hierarchy of aspect is needed, which we introduce **restaurant placement ranking model via leveraging user-generated contents (RESPECT)** which includes LDA extracted aspects, pre-defined aspects and a processed extracted aspects to capture more intensive information in aspect level. Also, the sentiment of the review is highly valued in social media and marketing fields, other than adding up the sentiment score and

viewing the performance as a whole, we can combine the sentiment score with aspect importance might create a new way to evaluate performance of the restaurant in an aspect level. Other than sentiment information, the review can also disclose the competitor relations, such as the degree of competitors clustering, with these novel features mentioned above may offer a new perspective to tackle this problem.

As for defining new store location selection, most of the work focuses on predicting the ratings or revenue of one restaurant to represent the success of opening in one area. By providing rankings of location candidates, not only simulates the real-world selection process, but reveals additional information to the decision maker. Therefore, this study develops a ranking model based on geo-economics-based features, location-based features, and social-media-based features which generates additional useful features to find the best restaurant location. We examine the following research questions in this research:

- Can sentiment analysis help to generate more useful features?
- Does the aspect retrieved from different ways affect performance?
- Apart from sentiment analysis, can the degree of competitors feature extracted from the review contribute to restaurant placement?

Chapter 2 Literature Review

2.1 Existing New Store Location Selection Methods

When dealing with the new store selection problem, most of the adopted features were mainly used to characterize the location and the store. The data source of the feature groups can be divided into three parts as mentioned before, namely: geo-economics-based features, location-based features, and social media-based features. In recent years, the introduction of Geographic Information Systems (GIS), location-based applications like map search, navigation systems and social media applications into the world has led to an increase of relevant studies employing location-based features, for example, point of interest (POI) check-in data and social-media-based features like user-generated contents (UGC). As can be seen from previous studies, different feature sources have been adopted simultaneously, for example, there has been a gradual shift from the use of only geo-economics-based features to the use of a mixture of the above features. Though sources changed throughout the year, the same concepts are still being captured. Feature groups for location, the overall concept is to describe whether the chosen location is suitable to withhold a new store opening, including demand, supply, traffic accessibility and the consumption power of the location. With regard to feature groups for stores, more focus is on capturing the performance and the competitiveness of the store itself, for example, willingness to pay of the customers, ratings, profit, etc.

2.1.1 Feature Groups for Location

Supply of an area is usually depicted by shop density and demand uses human flow or population in that area. Most works of the geo-economics-based features focus on, census data such as: population, birth rate, mortality rate, crime rate, immigration rate, replacement rate, etc. These features allow us to get a picture of that neighborhood's

supply and demand in a more quantify level. Location-based features such as check-in data and web search data provide insights of human flow, with the knowledge of these foot traffic, it is easier to choose the hot spot of the neighborhood. Karamshuk et al. (2013) used the dataset collected from Foursquare in New York. They managed to get features of geographic and mobility features that serve as demand for popularity prediction on three different coffee shops and restaurant chains (Starbucks, Dunkin' Donuts, and McDonald's) in New York City. Rahman and Nayeem (2017) also took advantage of Foursquare New York City to retrieve demand and the capability of gathering audiences and used it to predict a location for live campaigns. Xu et al. (2016) detects the gaps between demands and supply by extracting timestamps and locations through queries from the Baidu search engine. These extracted keywords are then transformed into a demand point to specify the temporal and spatial features of demands. Zhdanov et al. (2022) collected reviews from TripAdvisor to gain insight into restaurant distribution in Russia.

Most studies indicated that easy access to public transportation is an important factor that influences consumers to travel to the place. When it is more difficult to reach a destination, consumers are less likely to go there (Ge et al., 2019; Sokol & Jordanov, 2020; Yu et al., 2016). In most cases, traffic convenience is a necessary but not sufficient factor to the success of a shop. Some studies take distance to downtown and traffic accessibility as indicators to measure the transportation convenience (Li et al., 2021; Morrison & Abrahamse, 1996). Other studies uses maps and POI check-in data to capture the availability of public transportations (Guo et al., 2018; Rahman & Nayeem, 2017; Sokol & Jordanov, 2020; Zeng & Tang, 2019).

Morrison and Abrahamse (1996) takes account of future consumption base of a neighborhood, and specific demographic factors to the best location for a large

supermarket catering to one-stop shoppers. Geo-economics-based features takes household income, house price to represent the consumption power of a certain location, while location-based and social-media-based features offer valuable information about a user's geographical context interests, preferences, and social connections, they generally do not directly reveal information about a user's financial status, income, or wealth.

2.1.2 Feautures Groups for Stores

Competitiveness and complementary are two main features that most studies focus on retrieving. Geo-economic-based features take shop commercial profile, expertise insights to measure own competitiveness (Morrison & Abrahamse, 1996), while types created by user can be found in the source of location-based and social-media-based. These types can also be an input to calculate shop diversity. A handful of studies apply neighborhood entropy measure proposed by Karamshuk et al. (2013) by using location-based features (Sokol & Jordanov, 2020; Yu et al., 2016; Zeng & Tang, 2019). Additionally, social-media-based features, to be precise, UGCs, offers compatibility to an location also account for measure competitiveness. Wang et al. (2016) utilizes the comments by using LDA model to evaluate restaurants at an aspect level and calculate the compatibility with the location.

The competition or cooperation between shops in the neighborhood may greatly affect the success rate of the new entering shop. By measuring the network of commercial interactions between retail shops, and the network of interactions between the different commercial categories, Jensen Quality index (Jensen, 2006) and is widely used in calculating interactions between venues. Geo-economics-based features uses commercial profile to get the index, while location-based features and social media uses shop types to perform this calculation. Co-occurrence of venues in an area is also viewed as complementary, the higher frequent different store locating at the same location might

imply a shared advantage cooperation, Yu et al. (2016) models the complimentary and competition between shops, and Ahedo et al. (2021) models the balanced tenancy in an area, using location-based features to compute. Social-media-based features are also used, Wang et al. (2016) uses review to retrieve aspects as previously stated to calculate complementary.

2.1.3 Summary of Existing New Store Location Selection Methods

Below Table 2-1 and Table 2-2 is a summary of location feature groups and store feature groups used by studies. By looking at the table we can observe that the source type of features has become hybrid instead of just using a single source type, meaning more data are take into thought patching together a holistic view of the store or location. We can also observe that both location-based and social-media-based features contains more information in a single data; hence, less source is needed to construct the feature groups. Another worth noting is that for location feature groups supply and demand is the most captured feature and both competitiveness and complementary for characterizing store details.

Table 2-1 Summary of Feature Groups for Location

Features			**Studies**									
Feature Groups	Source	Source Type	Morrison and Abrahamse (1996)	Karamshuk et al. (2013)	Xu et al. (2016)	Wang et al. (2016)	Rahman and Nayeem (2017)	Guo et al. (2018)	Zeng and Tang (2019)	Sokol and Jordanov (2020)	Li et al. (2021)	Zhdanov et al. (2022)
Supply	Commercial profile	Geo-economics	V	V	V							
	POI check-in	Location		V			V	V	V	V	V	V
	UGC	Social Media			V	V					V	V
Demand	Census data	Geo-economics	V		V	V						
	POI check-in	Location		V			V	V	V	V	V	V
	web search				V							
	UGC	Social Media			V						V	V
Transportation	Traffic accessibility	Geo-economics	V		V							
	Distance to downtown				V							
	POI check-in	Location					V	V	V	V		
Consumption Power	Household income	Geo-economics	V		V		V					
	House price							V				

Table 2-2 Summary of Feature Groups for Store

Features			Studies									
Feature Groups	Source	Source Type	Morrison and Abrahamse (1996)	Karamshuk et al. (2013)	Xu et al. (2016)	Wang et al. (2016)	Rahman and Nayeem (2017)	Guo et al. (2018)	Zeng and Tang (2019)	Sokol and Jordanov (2020)	Li et al. (2021)	Zhdanov et al. (2022)
Competitiveness	Commercial profile	Geo-economics	V		V			V				
	Expertise insights		V									
	POI check-in	Location			V	V			V	V	V	V
	UGC	Social Media				V					V	V
Complementary	Commercial Profile	Geo-economics		V	V	V		V		V	V	
	POI check-in	Location		V				V		V	V	V
	UGC	Social Media				V						

2.2 Prediction tasks

Prior studies define the new store location selection problem into two tasks, namely, performance prediction and ranking. Performance prediction task focus on predicting one restaurant or store's performance at a time, while ranking takes other candidates into consideration and return a ranked list. Whether its performance predicting or ranking, the success of a store is mostly valued by popularity, thus, check-in number is an important source serve as the dependent variable. Also, as mentioned before ranking task were rarely used in store location selection problems, most of the works still view the problem as a predicting task. Various studies of store site selection are organized below in Table 2-3 distinguished by the selection of the predicted value.

Table 2-3 Summary of Prior Prediction Task

Problem	Label(y)	Studies (years)
Prediction	Revenue	Morrison and Abrahamse (1996); Li et al. (2021)
	Check-in Count	Karamshuk et al. (2013); Wang et al. (2016) ; Zeng and Tang (2019); Sokol and Jordanov (2020)
	Ratings	Guo et al. (2018)
	Post Popularity	Li et al. (2021)
Ranking	Check-in Count	Xu et al. (2016); Rahman and Nayeem (2017)
	Ratings + Review Count + Review Sentiment	Zhdanov et al. (2022)
	Revenue	Damavandi et al. (2019)

2.3 Summary of Literature Review

To sum up, most works apply geo-economics-based feature and location-based features, which are restaurant-independent. They only reflect the general area properties of the locations without taking the restaurant properties into consideration. Social-media-based features generate attributes unique to the restaurant itself, depicting restaurants in a more personal view, but most work stops at the phase of rating and review count, instead of using the content itself. Whether its prediction or ranking task, check-in count or review count are the most commonly used value to show the successfulness of one restaurant, hence, we follow this method to design our dependent variable.

Current research gaps are as follow:

- Few works focus on the content of review to extract competitiveness features, ignoring the details of certain aspects when evaluating a restaurant.

- Using LDA model to do aspect extraction results in lack of interpretability, the generated topics may not directly correspond to meaningful or intuitive aspects that users are interested in. And since LDA assumes that each word in a document is associated with a single aspect, this may lead to difficulty in handling overlapping aspects causing the model to assign words to separate aspects even when they are closely related.
- Sentiment analysis are applied in a general level by summing up the score to represent the restaurant, blurring the strength and weakness of a restaurant in a certain aspect.

The novelty methods of our research are:

- Instead of using LDA to do aspect extraction, we also promote two techniques to extract more precise aspects and evaluate a restaurant's performance in a more comprehensive approach.
- Aspect-based sentiment analysis are applied and calculated as indicators of measuring the compatibility of the restaurant and the location.
- Besides using restaurant types to represent the competitors, we deploy community detection to cluster the competitors, creating a novel feature for the task.

Below Table 2-4 and Table 2-5 two tables are presented, showcasing the features employed in our method and a comparison with prior studies. The majority of the factors utilized in our approach are social-media-based features, with our primary focus centered on the assessment of demand and supply for location-based feature groups, aligning closely with recent research efforts in this domain.

Table 2-4 Our Method Compared with Prior Works (Feature Group for Location)

Features			**Studies**										
Feature Groups	Source	Source Type	Morrison and Abrahamse (1996)	Karamshuk et al. (2013)	Xu et al. (2016)	Wang et al. (2016)	Rahman and Nayeem (2017)	Guo et al. (2018)	Zeng and Tang (2019)	Sokol and Jordanov (2020)	Li et al. (2021)	Zhdanov et al. (2022)	**Our method (2023)**
Supply	Commercial profile	Geo-economics	V	V	V								**V**
	POI check-in	Location		V			V	V	V	V	V	V	
	UGC	Social Media			V	V					V	V	
Demand	Census data	Geo-economics	V		V	V							
	POI check-in	Location		V			V	V	V	V	V	V	**V**
	web search				V								
	UGC	Social Media			V						V	V	
Transportation	Traffic accessibility	Geo-economics	V		V								
	Distance to downtown				V								
	POI check-in	Location					V	V	V	V			
Consumption Power	Household income	Geo-economics	V		V		V						
	House price							V					

Table 2-5 Our Method Compared with Prior Works (Feature Group for Store)

Features			**Studies**										
Feature Groups	Source	Source Type	Morrison and Abrahamse (1996)	Karamshuk et al. (2013)	Xu et al. (2016)	Wang et al. (2016)	Rahman and Nayeem (2017)	Guo et al. (2018)	Zeng and Tang (2019)	Sokol and Jordanov (2020)	Li et al. (2021)	Zhdanov et al. (2022)	**Our method (2023)**
Competitiveness	Commer-cial profile	Geo-economics	V		V			V					
	Expertise insights		V										
	POI check-in	Location			V	V			V	V	V	V	
	UGC	Social Media				V					V	V	**V**
Complementary	Commer-cial Profile	Geo-economics		V	V	V		V		V	V		
	POI check-in	Location		V				V		V	V	V	
	UGC	Social Media				V							

Chapter 3 Our Proposed Method: RESPECT

3.1 Problem Formulation

We define the ranking task as follows in our study, given a set of reviews and location-based attributes we create features including location-based features, aspect-based features and competitor-based features. We define our problem as a ranking problem, the goal is to have the predicted ranking be as close to the true rank list as possible for a brand with n new stores to establish, given $m\,(m >> n)$ candidate locations to consider. We exploit different ranking methods including pointwise and listwise to get the best ranking result. Pointwise prediction is considered as a binary classification, while listwise ranking generates a list of ranked locations.

3.2 Overview of Our Proposed Architecture: RESPECT

We define the new store selection problem as a ranking task and propose our method as a restaurant placement ranking model via leveraging user-generated contents. Data collected from the Yelp dataset are then transformed into three types of features: geo-economics-based features, location-based features and social-media-based features. Social-media-based features are being processed through abstract-based sentiment analysis (ABSA) and community detection to learn any hidden information from user generated contents and to examine restaurants in a finer measure. Lastly, these features are trained by ranking model to predict the rank of a list of candidate location to open a new branch for an existing brand of restaurants. The model structure we proposed is presented in below Figure 3-1.

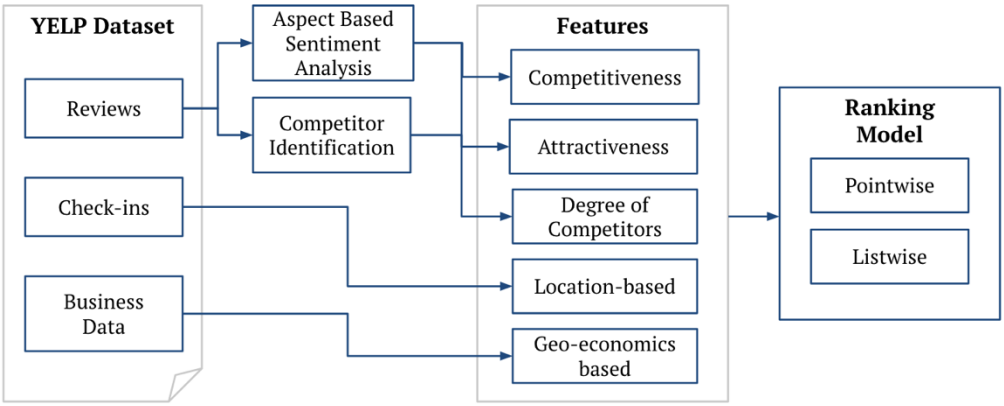


Figure 3-1 Structure of RESPECT

3.3 Geo-economics and Location Based Features

We exploit geo-economics-based features and location-based features to capture location and store feature groups inspired by prior works. In our study, we extract these features from the Yelp dataset which includes commercial profiles, check-in data and user-generated contents including ratings, comments and photos.

Density which accounts for showing people how concentrate one location is with restaurants, where total number of restaurant located in l is presented as $N(l)$:

$$x_{vl} = N(l) \tag{3.1}$$

The density of one location not only shows the supply but also tells the opportunity or threat in that place in a general way. The higher density in one place shows that the more restaurant is located in that area, hence a higher chance to lure people in, but on the downside meaning there would be an intense competition.

Area popularity shows the visitors amount in one location, where $C(l)$ represents the check-in set located in location l, area popularity is through aggregating the number of check-in count $C(l)$ located in location l :

$$x_{cl} = |\{c \in C(l)\}| \tag{3.2}$$

The area's popularity serves as a significant indicator of the demand, playing a critical role in determining the prospects of a restaurant's success. A higher level of popularity not only signifies a greater number of potential customers but also indicates a thriving and bustling environment that can attract a diverse range of patrons. This agglomeration effect, characterized by the concentration of businesses and customers in a specific location, creates a vibrant and competitive atmosphere that fosters synergies and mutual benefits for establishments in the vicinity.

Neighborhood Entropy is used to measure the diversity of restaurants in one area, calculated by using entropy technique and sum up the frequencies of the different categories of restaurant in the location. The category frequency is calculated as follow, with the category set Γ, the number of restaurants that has the same category γ as restaurant v denotes as $N_\gamma(l)$ is divide by the total number of restaurants $N(l)$ located in l:

$$x_{vl} = -\sum_{\gamma \in \Gamma} \frac{N_\gamma(l)}{N(l)} * log \frac{N_\gamma(l)}{N(l)} \tag{3.3}$$

Neighborhood entropy shows the variety of restaurant options in the location, the higher value the more heterogeneous, vice versa. If the location has a low value of neighborhood entropy, it may be pointing out that the location has a preference towards a certain type of restaurant, sending a signal the restaurant with the same type this might be a good place to start or to open a new branch.

3.4 ABSA for Social-Media-Based Features

User-generated contents occupy a large part of social-media-based features, and since it contains an abundant amount of reviews for just about anything, they serve as an important source of measuring the performance or showing preference towards the commented objects, including store or restaurants. Wang et al. (2016) uses the LDA

model to retrieve aspects that serve as criteria to evaluate a restaurant in multiple facets. The work opens a way for us to measure stores or restaurants in a fine-grained metrics, following his work, we use different tactics to get aspects and recreate features to represent the competitiveness and performance of the targeted restaurant. With the help of aspect extraction sentiment analysis can also be leveled up to aspect scope, instead of just talking about the overall performance, we can now analyze the restaurant in different attributes, finding deeper strength and weakness within restaurants.

Our proposed method RESPECT is based on those aspect extraction to create novel features, the methods are as follow: LDA model aspect extraction, pre-defined aspects and pretrained model aspect extraction. Novel features are generated to represent performance of the restaurant and its compatibility with the targeted locations, in other words the competitiveness of the restaurant the specified location. In the following section, we will introduce the mechanism used to extract aspects and how the features are being captured.

3.4.1 Latent Dirichlet Allocation Extracted Aspects

Latent Dirichlet Allocation (LDA) is a probabilistic generative model commonly used for topic modeling. It assumes that documents are represented as a mixture of topics, and each topic is characterized by a distribution of words. Given a set of restaurants V, restaurant v where $v \in V$, the reviews gathered from restaurant v is viewed as a single document d. After training the model with all the documents, we can get the probability of a document belonging to a topic k, written as $t_{vk} = p(k|v)$. The goal of LDA is to infer the underlying topic structure of a collection of documents. It works by iteratively assigning words in the documents to different topics and adjusting the topic-word distributions to maximize the likelihood of the observed data, assuming the words in a document are generated based on a probabilistic process involving the topics and their

associated word distributions. Before training the document, reviews will go through a series of text processing. First, we tokenize the sentences by spacebar, then we do POS tagging and lemmatize the tokens. After that we remove English stop words and numbers, leaving ‘nouns’ only. Next, tokens occurred more than 50% of documents and less than 50 documents are also ignored. Lastly, top 1000 most frequent tokens are kept for LDA training. Once the LDA model is trained, you can access the inferred topic distributions for each document then the aspects are then retrieved using the most important word of the topic, every word represents one topic only. The retrieved 20 aspects are as follows: 'burger, table, beer, taco, order, cream, steak, drink, pizza, sandwich, server, breakfast, wing, sauce, room, chicken, dish, rice, roll, salad'

3.4.2 Pre-defined Aspects

After observing the important words generated by the LDA model for every topic, we can see that those “topics” aren’t exactly what we called “aspects”. Hence, we carefully chose 30 words as our aspect for analyzing the sentiment of the reviews. The pre-defined aspects are chosen from the Michelin guide and some general assessment indicators in a heuristic matter, which are not only used by inspectors around the world to give out unbiased evaluations to the restaurants, but also critics and reviewers around the world. A total of 30 words are chosen regarding six dimensions, namely, Food Quality, Service, Environment, Value for Money, Cleanliness, Ambiance, Menu Variety, Portion Size, Presentation, Staff Friendliness, Speed of Service, Accuracy of Orders, Dietary Options, Drink Quality, Reservation Process, Noise Level, Location, Overall Experience, Authenticity, Innovation, Beverage Selection, Wine List, Cocktail Quality, Dessert Quality, Seating Comfort, Outdoor Seating, Special Occasion Suitability, Child-Friendliness, Parking Availability, Online Ordering Experience.

3.4.3 Pretrained Model Extraction

Aspect term extraction (ATE) is an ABSA subtask, which includes: ATE, ATSC, and Joint Task, that involves identifying and extracting specific aspects or entities of interest from text data. The goal of aspect extraction is to determine the key topics or features being discussed within the text, enabling a more fine-grained analysis of sentiment or opinion towards those aspects. InstructABSA (Scaria et al., 2023) is a pretrained model we use for extracting aspects in our experiment. The model was trained using the SemEval (Semantic Evaluation) dataset (SemEval 2014 Laptops and Restaurants) and has surpassed previous SOTA on downstream ABSA subtasks by significant margin. The model we are using is the ATE model. It returns the aspects in a sentence if there are any in the sentence, if not it returns "noaspectterm". After gaining total of 209,584 extracted terms from review, we did some filtering and grouping to make the aspects more general. The taken steps are as follow:

I. lemmatized and remove words that are empty or has 'restaurant', 'naan' and 'noaspectterm' in it.

II. Since most the words have document frequency (df) less than 10 with standard deviation of 100, we filter the aspects by df < 100

III. Then group the words following the former dimensions into 'food', 'drink', 'menu variety', 'price', 'ambience', 'service' and 'others', words that belong to 'others' are left the same. After processing the aspects, total of **180 aspects** are captured.

Table 3-1 depicts the number of aspects at each stage of the grouping process is presented, along with the average document frequency, illustrating the diversity of the extracted aspects. Furthermore, the results of decreasing average document frequency indicate that after undergoing the grouping process, these aspects become more generalized, alleviating their initial granularity.

Table 3-1 Summary of Extracted Aspects

	Original	Lemmatized	Filtered	Grouped
Aspect Count	209,584	148,610	944	180
Average document frequency	7.13	9.72	1154.15	4173.29

Examples of using the pretrained model is down below in Table 3-2:

Table 3-2 Example of extracted aspects from InstructABSA

	Original Text	Extracted Aspects
Example 1	Quite simply, THE go-to restaurant in Clearwater Beach. Try the **grouper sandwich** (Buffalo style) and enjoy their **drink specials**!	[‘grouper sandwich’, ‘drink specials’]
Example 2	We just had the most amazing **meal**! The **environment**, **food**, **staff** were exceptional! We were a large group on a Friday night and it couldn't have been better!	[‘meal’, ‘environment’, ‘food’, ‘staff’]
Example 3	The **Sicilian slice**, fresh out of the oven, was one of the most perfect slices of **pizza** I've had in America. The man who runs this place is a Spanish-speaking Italian grandfather. I want the calzones for birthdays, Christmas and Hanukkah.	[‘Sicilian slice’, ‘pizza’]

Feature Selection are implemented on the competitiveness features constructed by this type of retrieved aspects, since, through grouping and frequent filtering there are still 180 aspects left. Comparing with the number of aspects generated from former methods, it is significantly larger than both of them almost 5-6 times more. Therefore, we would like to reduce the features to avoid them becoming noises to our model. We

implement feature selection through ranking the importance of feature created on the basis of these aspects by chi-square test. The chi-square test is especially useful when dealing with categorical variables and a categorical target variable. It helps identify features that have a significant impact on the target variable and can be used to reduce the dimensionality of the dataset, leading to more efficient and potentially more accurate models. We calculate the chi-square score of the competitiveness features between two groups, namely locations that the stores actually open there, and locations that store did not open. Locations are marked as 1 if the store really locates at the place and 0 for the candidates. After the chi-square test, we choose top 25% of the aspect constructed competitiveness feature as our model input.

3.4.4 Sentiment Analysis

Sentiment analysis is the process of determining the sentiment expressed in a piece of text, such as a review, social media post, or customer feedback. It involves analyzing and classifying the subjective information present in the text to identify the underlying sentiment or attitude conveyed. PyABSA (Yang & Li, 2022) is a pre-trained model we use for the sentiment analysis. Given an aspect, it outputs the probability of the sentiment as follow: negative, neutral and positive. The probabilities of each sentiment is between 0 and 1. After retrieving the aspects from last step, we then throw a review which consists of multiple sentences, and a aspect as a pair into the pretrained model and get the sentiment probabilities. A restaurant may have multiple reviews, and the reviews may contain the same aspects. The sentiment of the same aspects calculated by averaging the scores, so that unique aspects and sentiment scores are used to represent a restaurant. Examples of using the pretrained model is shown in Table 3-3 down below:

Table 3-3 Example of Sentiment Analysis

ABSA Method	Original Text	Extracted Aspects	Predicted Sentiment
Example 1- LDA	Quite simply, THE go-to restaurant in Clearwater Beach. Try the **grouper sandwich** (Buffalo style) and enjoy their **drink specials**!	burger	pos
		table	neu
		beer	pos
Example 2- Pre-defined	We dared the food even though it was busy. First we had fish tacos, they were good. Those came out of the kitchen within five minutes. Then we ordered the loaded tator tots. They forgot about our order and we had to wait over 45 minutes. Great though, I would order them again. Prices too high. Vodka mixed drink was $11, double amaretto sours were $20 each. The beer bucket was $30 for 6 beers.	value for money	neg
		service	neg
		food quality	pos
Example 3- InstructABSA	We stopped in for a quick bite while walking the boardwalk and were not disappointed. There is not a ton of seating, so we had a wait a few minutes,	seating	neg
		table	neu

	but they moved quickly, cleared a table, and got us set up to order. I had the fish tacos, and they were excellent - I would go back to have them again. I also had a taste of the lobster roll and it was amazing - definitely will order that if I'm ever back in town. All in all, this place is somewhat of a hidden gem you don't want to miss!	lobster roll	pos
		wait	pos

3.4.5 Social-media-based features construction

After gaining the extracted aspects through multiple channel, we construct them into following inputs to capture certain hidden information, we also implemented the features that were performed in Wang et al. (2016) to represent attractiveness of the restaurants, since it is the first to create a score in aspect level to evaluate the compatibility of a restaurant and the location. Both attractiveness features utilizes the probability of topics given by the LDA model, which act as a aspect for evaluation of the restaurant. Continuing along the probabilities gain from the LDA model, the aspect k probability for a restaurant v can be represented as t_{vk}, $\theta_v = \{t_{v1}, t_{v2}, t_{v3}, \dots, t_{vk} | 0 \leq t_{vk} \leq 1, and \sum_{k=1}^{K} t_{vk} = 1\}$ is the set of probabilities for all the aspect extracted from LDA. The probability of aspect k at a location is the combination of all the restaurants located in location l, which is presented as $t_{lk} = \frac{1}{Z}\sum_{v \epsilon A(l)} \log(N(v) + 1) t_{vk}$. With these 2 features in hand, we can construct the attractiveness features.

Affinity measures the harmony between consumers of restaurant v and those at location l. The formulation is as follow:

$$x_{vl} = \{t_{vk} * t_{lk}, 1 \leq k \leq K\} \quad (3.4)$$

The more similar preference of customer in restaurant v and those in location l, the higher the score, vice versa. If the score is high, it is safe to say that it will be easier to predict the success rate in a new place, since you can observe your customers and foresee the reaction of the consumer at location l.

Complementary measures the supplementary between consumers of restaurant v and those at location l, which means measuring the opposite preference between them.

$$x_{vl} = \{t_{vk} * (1 - t_{lk}), 1 \leq k \leq K\} \quad (3.5)$$

The score obtained from analyzing the behavior of the two distinct groups of consumers reveals their contrasting preferences and tendencies, providing valuable insights into their distinct patterns of interaction with businesses and offerings. This disparity in behavior can be viewed as a potential opportunity or a potential threat, contingent on the receptiveness and adaptability of the consumer base in a particular location. Then it comes down to the openness of the consumer in location l. In scenarios where consumers in location l exhibit a sense of openness, curiosity, and a propensity to explore novel experiences and offerings, the observed behavioral differences can be regarded as an opportunity. Conversely, when consumers at location l display a more conservative or resistant attitude towards novel offerings, the identified behavioral contrasts may pose a threat. Failing to align with the prevailing tastes and preferences could result in reduced consumer engagement and limited market acceptance, hindering business growth and expansion in that specific location.

Competitiveness features which include two features both compose of two sub features. First is **importance**, we measure importance by the number of the aspect was mentioned through all the reviews. $Importance_{vk}$ refers to the mention time that aspect k being mentioned in restaurant v's review and $Importance_{lk}$ refers to the mention

time that aspect k being mentioned in the reviews of all the restaurants at location l. Next, $Performance\_gap_{vl,k}$, just as it sounds, it measures how well the restaurant performs comparing to the other restaurant underlying at location l, by subtracting the sentiment score of aspect k under restaurant v's review with the sentiment score of aspect k under all the reviews from restaurants locating at location l.

Strength is measure by multiplying importance of restaurant v to the performance gap.

$$x_{vl} = \left\{ Importance_{vk} * Performance_{gap_{vl,k}}, 1 \leq k \leq K \right\} \quad (3.6)$$

The higher the score meaning the restaurants outperforms on certain aspect that the consumers of restaurant v care about. Since you are satisfying your consumers a lot, it becomes a testament to the restaurant's strengths, affirming its ability to meet and even exceed the expectations and preferences of its loyal customer base. These satisfied customers, in turn, contribute to the restaurant's positive reputation, word-of-mouth referrals, and repeat business, further solidifying its position as a preferred choice among the local dining options. In the competitive landscape of the restaurant industry, the restaurant's ability to consistently deliver exceptional experiences and cater to its consumers' desires grants it a competitive edge over rival establishments. This strength positions the restaurant to attract new customers in the new location, thereby bolstering its market presence and augmenting its overall performance.

Opportunity is measure by multiplying importance of location l to the performance gap.

$$x_{vl} = \left\{ Importance_{lk} * Performance_{gap_{vl,k}}, 1 \leq k \leq K \right\} \quad (3.7)$$

The elevated score demonstrates the restaurant's ability to meet and exceed the expectations of its target customers in location l, thereby fostering a positive reputation and building strong customer loyalty. This competitive advantage positions the restaurant

as a preferred choice among the local dining options, attracting a steady stream of patrons and ensuring consistent business success. Given the restaurant's exceptional performance and favorable reception in location l, venturing into opening a new branch or store becomes a logical and strategic move. Capitalizing on the trust and goodwill it has garnered, the restaurant can leverage its successful track record to appeal to a broader customer base and expand its market presence.

3.5 Competitor Identification for Social-Media-Based Feature

In our study, we implement competitor identification through community detection. Community detection refers to the process of identifying groups or communities of reviewers who exhibit similar patterns, preferences, or behaviors. It involves analyzing the relationships or interactions among reviewers based on their reviews to uncover meaningful clusters or communities. The graph that we implement community detection is constructed through the relation of 'co-mention'. We construct the co-mention matrix with the sampled documents we used earlier for LDA training to avoid getting reviews from the predicted areas. The co-mention matrix is undirected, each entry in the matrix corresponds to the count of co-mentions between two restaurants. Once the matrix is constructed, we also reate a weighted graph using Jaccard similarity coefficient. It compares the overlap between the sets of nodes in different communities, providing a measure of how similar two communities are based on their shared nodes. Then, we use modularity-based approach to get the best partition and clustering of all the restaurants.

3.5.1 Example of the identified competitor groups

Restaurants are being grouped into 10 groups, here are some examples of the groups. After looking through the grouped restaurants in each group we can sort of classify these group into bellow categories (the bolded restaurants are those who related to the group):

- **Group 1 - Mexican Food**

  ‘Sauce on the Side’, ‘El Diablo Burritos’, ‘Las Palmas Mexican Restaurant’, ‘Chimi's Fresh-Mex’, ‘California Tortilla’, ‘Grand Hacienda’, ’Bisbee Breakfast Club’, ‘Black Bear Diner’

- **Group 2 – Grills and steak**

  ‘Cheddar's Scratch Kitchen’, 'Uno Pizzeria & Grill', ‘Ruth's Chris Steak House’, ‘Rubio's’, ‘Nifty Fifty's’

- **Group 3 – American Style**

  ‘Burger Republic’, ‘Dave & Buster's’, ‘Gyro Shack’, ‘Blake's Lotaburger’, ‘Burger Monger’, ‘Village Inn’, ‘Joey's New York Pizza and Italian Restaurant’, ‘SoFresh’, ‘Boise Juice Company’, ‘Token Ramen & Kung Fu Tea’, ‘Irish 31’, ‘Granite City Food & Brewery’

We can see that by gather clusters through co-mentions, these restaurants do display similar categories or offer the same type of meal. Some may seem unrelated to the category type, it might be that the co-mentions are geographical related meaning that the restaurant are in the same area so it is mentioned, or there are some underlying facts that the consumer tend to assume them as competitor or comparative sample.

3.5.2 Feature Construction

Degree of Competitors measures the competitor number of restaurant v in location l. Restaurants are count as restaurant v’s competitor if they are clustered into the same group $G(v)$. Earlier, we have created a Jaccard Matrix using the co-mention matrix, we add up the Jaccard weight if the restaurants are in the same group with restaurant v.

$$x_{vl} = \sum_{t=1}^{V} w_t \ , \begin{cases} w_t, if\ v_t \in G(v) \\ 0 \end{cases} , t \in V \tag{3.8}$$

As the degree of competitors increases, it indicates a higher level of competition in the chosen location, with a greater number of competitors operating within the same area. Value of the degree of competitors reflects a densely saturated market environment, where numerous restaurants battle for the attention and patronage of the local consumer base. This intensified competition can pose both challenges and opportunities for businesses seeking to establish a presence in the area.

Chapter 4 Empirical Evaluation

4.1 Dataset Collection

Dataset is offered by Yelp[7], containing businesses, reviews, and user data for use in personal, educational, and academic purposes. It offers over 6,990,280 reviews, 150,346 businesses, 200,100 pictures and 11 metropolitan areas by 1,987,897 users. Since we only focus on restaurants in our study, business without 'restaurant' or 'food' in their category are excluded. Date are narrowed down to 2017-01-01 ~ 2022-12-31. Restaurants that have 3 -14 branches and still open till 2022-12-31 with at least 20 reviews are chosen. Table 4-1 shows the statistics of our dataset.

Table 4-1 Statistics of Adopted Restaurant Dataset

Statistics on Yelp Dataset	
# Restaurants / branches	401 / 1848
# Locations	407
# Reviews	387,723
# Avg.. Reviews	966
# Restaurant categories	880

4.2 Evaluation Procedure and Metrics

4.2.1 Dataset Construction

We devised our train test set construction process by partitioning the branches under the same restaurant into a 2:1 ratio for each restaurant in our data collection. Following this separation, we paired these branches with negative samples representing locations

[7] Yelp Open Dataset 2023 Available at https://www.yelp.com/dataset (Retrieved on Jan 2022).

where the restaurant did not open a store. For every restaurant, a total of 20 locations were carefully selected based on their postal code, we assign the negative sample in the following order, the highest-ranked branch by review count, followed by the second-highest, and so on.

Regarding the attributes representing both locations and restaurants, we ensured consistency within each restaurant's group by using the attributes of the testing group to represent the training group and vice versa. For every location representation concerning a specific restaurant *v*, the location was represented by all other restaurants that operate there, excluding restaurant *v*.

4.2.2 Evaluation Metrics

In order to have unbiased and reliable result, bootstrapping is executed for every experiment and the score we present is averaged among the result. Below is the metric we use to evaluated our models, precision@1, precision@3, mean reciprocal rank (MRR) and normalized discounted cumulative gain (NDCG).

$$p@k = \frac{\#\ of\ recommended\ items\ @k\ that\ are\ relevant}{\#\ of\ recommended\ items\ @k} \tag{4.1}$$

$$MRR = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{4.2}$$

$$NDCG_k = \frac{DCG_k}{IDCG_k},$$
$$DCG_k = \sum_{i=1}^{k}\frac{rel_i}{log_2(i+1)}, \tag{4.3}$$
$$IDCG_k = \sum_{i=1}^{|REL_k|}\frac{rel_i}{log_2(i+1)}$$

4.2.3 Training Document Selection

To lower the computation power and time, we randomly picked ⅓ of reviews from

every restaurant, to simulate the unseen docs for every branch, and 20 topics were chosen. This review is used for LDA model training and community detection. We believe that by picking ⅓ of the review and with train test splitting by ⅓ , chances are relatively small peeking through the already established store or branch.

4.3 Experimental Setup

4.3.1 Implementation Details

In the context of Latent Dirichlet Allocation (LDA), perplexity and coherence are two commonly used measures to evaluate the quality and performance of a topic model. Perplexity is a measure of how well a trained LDA model predicts a held-out or unseen dataset. It quantifies how surprised the model is when trying to predict new documents. A lower perplexity score indicates that the model is better at predicting unseen data.

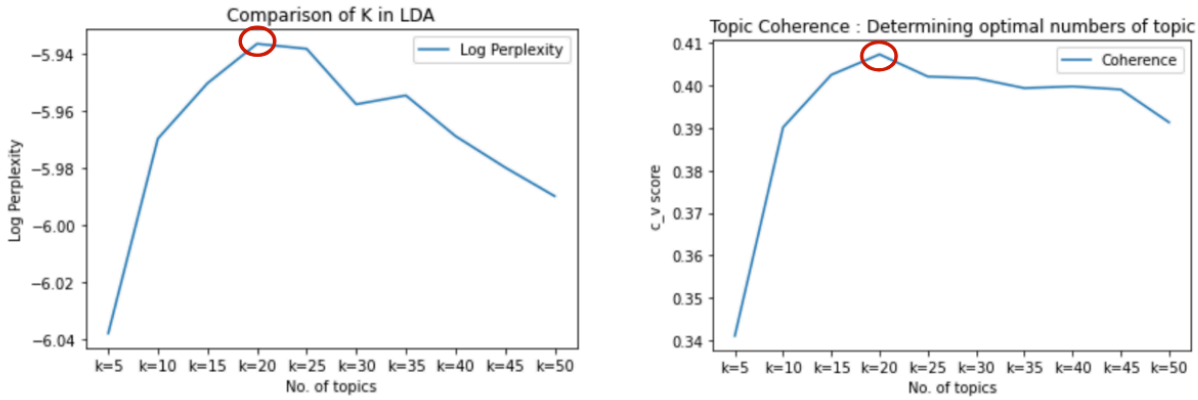


Figure 4-1 Perplexity and Coherence for Different Topic Number

Coherence is a measure of how interpretable and meaningful the topics generated by an LDA model are. It assesses the semantic coherence between words within each topic. Higher coherence values indicate that the topics are more coherent and understandable. Coherence measures typically consider the pairwise similarity between words based on their co-occurrence patterns within a given corpus. Both measures are used to determine the number of topic for LDA model. Sampled document as prior mentioned is used perform LDA training and evaluation. After seeing the result Fig.3 20 topics are picked.

The Chi-square test of deciding how many aspects should be kept is evaluated through NDCG. We see the performance of the remaining aspects and decide how many percentages of them should be kept. Figure 4-2 below shows that 25% of the aspect being kept can form the highest score, hence for the following experiments we use top 25% of the aspects in our RESPECT(InstructABSA) model.

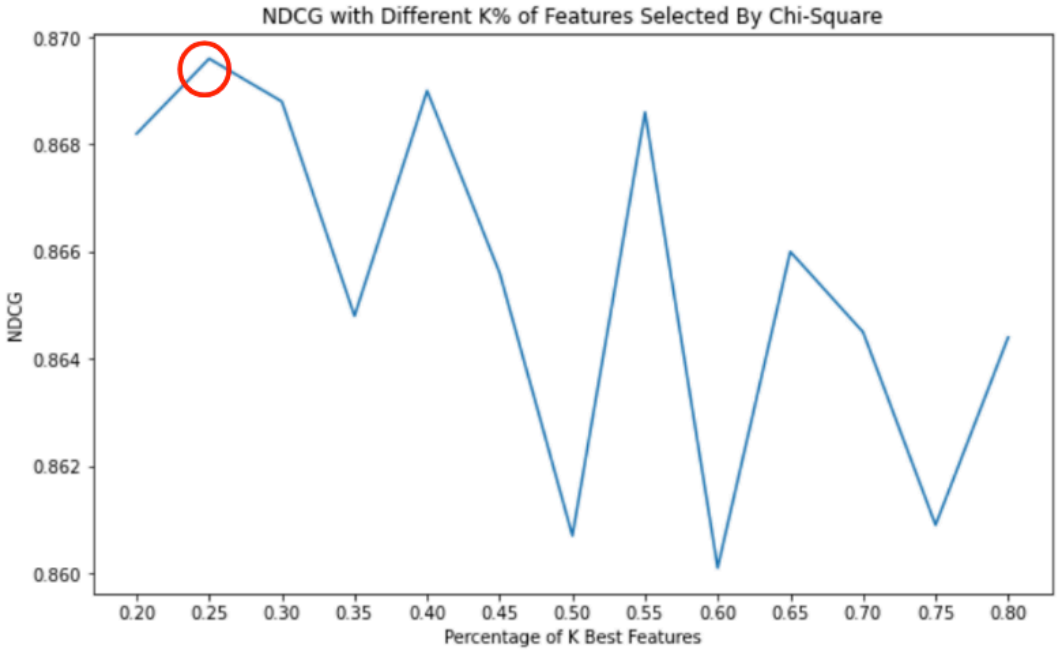


Figure 4-2 NDCG for Different Percentages of Aspects Kept

As for modularity approach, we chose the partition with the highest modularity score. The modularity score represents the quality of a given partitioning of nodes into communities within a graph. It quantifies the degree to which the graph's internal edge density is higher within communities compared to what would be expected in a random graph with the same degree distribution. It measures the difference between the actual number of edges between nodes in the same community and the expected number of edges based on a random graph model. It allows us to evaluate how well a given partitioning of the graph captures the community structure. A higher modularity score indicates a better-defined community structure, where nodes within each community are more densely connected to each other than to nodes in other communities. Shown in

below Figure 4-3, our best partition is 10.

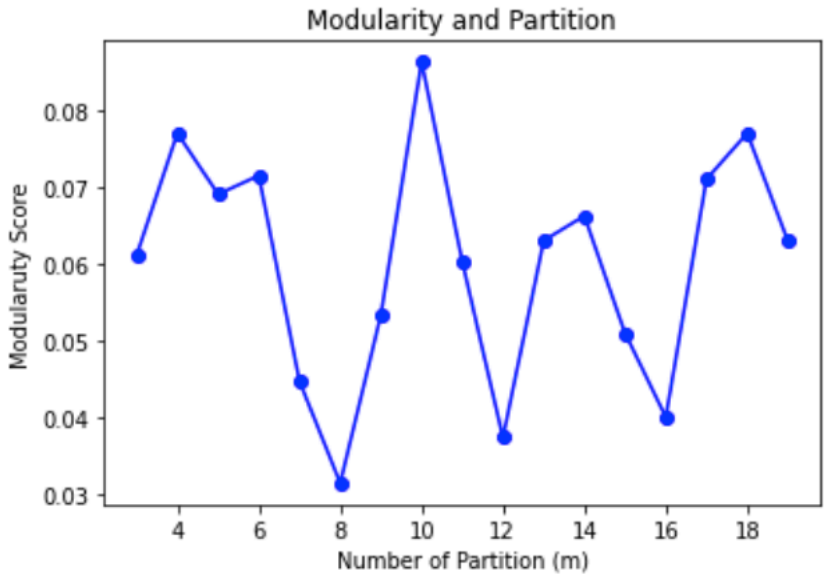


Figure 4-3 Modularity Score under Different Partition

And for our two ranking methods: pointwise and listwise. The data constructions are as follow. For pointwise ranking, two classification algorithms provided by Scikit-Learn: logistic regression (LR), random forest (RF), support vector classifier (SVC) and decision tree classifier (DTC). As for listwise ranking, LGBMRanker (Ke et al., 2017) a gradient boosting decision tree based learn to rank model is used. The relevance scores are given by sorting through review count of the branch, starting from 20 and subtract one in sequence.

4.3.2 Comparative Methods

We compare our model with following methods using different extracted features:

- Baseline: Geo-economics, location-based and attractiveness features
- RESPECT(LDA): Geo-economics, location-based, attractiveness features, LDA extracted competitiveness features and degree of competitor
- RESPECT(pre-defined): Geo-economics, location-based, attractiveness features, pre-defined competitiveness features and degree of competitor

- RESPECT(InstructABSA): Geo-economics, location-based, attractiveness features, InstructABSA extracted competitiveness features and degree of competitor

Below Table 4-2 shows the differences of baseline and our proposed method

Table 4-2 Summary of Features Used in Different Methods

Features Used	Baseline	RESPECT (LDA)	RESPECT (pre-defined)	RESPECT (InstructABSA)
Geo-economics-based	V	V	V	V
Location-based	V	V	V	V
Attractiveness	V	V	V	V
Competitiveness		V	V	V
Degree of Competitors		V	V	V

4.4 Evaluation Results

The result of our experiments is shown in Table 4-3.

Method	Model	precision@1	precision@3	MRR	NDCG
Baseline (Wang et al., 2016)	LR	6.52%	4.65%	22.71%	35.77%
	SVC	7.76%	4.55%	22.68%	35.23%
	RF	21.57%	14.60%	31.63%	45.60%
	DTC	21.82%	14.60%	31.67%	45.68%
RESPECT (LDA)	LR	6.36%	4.95%	23.14%	36.92%
	SVC	9.66%	7.11%	29.98%	44.28%
	RF	22.45%	17.32%	34.64%	50.00%
	DTC	21.95%	17.37%	37.06%	53.42%
RESPECT (pre- defined)	LR	8.67%	6.11%	23.51%	36.96%
	SVC	12.10%	8.06%	27.48%	40.60%
	RF	27.57%	20.75%	37.80%	54.68%
	DTC	28.82%	24.71%	41.00%	56.74%
RESPECT (InstructABSA)	LR	7.17%	5.02%	24.07%	36.84%
	SVC	9.85%	6.94%	29.75%	43.73%
	RF	22.95%	17.71%	35.05%	50.82%
	DTC	21.95%	17.42%	34.10%	49.97%

Table 4-3 Comparison Between Benchmark and RESPECT (Listwise Ranking)

Method	Model	precision@1	precision@3	MRR	NDCG
Baseline (Wang et al., 2016)	LGBMRanker	44.59%	36.31%	59.23%	69.11%
RESPECT (LDA)		58.12%	43.94%	70.20%	76.41%
RESPECT (pre- defined)		50.76%	39.33%	65.03%	73.11%
RESPECT (InstructABSA)		**74.86%**	**53.78%**	**83.18%**	**87.21%**

From the table above we can see that three methods of RESPECT outperform the baseline method. Pointing out that including sentiment in aspect level do create useful information since it contains not only preference towards the restaurant but also how well it performs while ranking with all the other restaurants, making the competitiveness customize to every restaurant itself. And a twist of competitor count, we group new clusters of competition groups by the review instead of making use of the categories provided through commercial profile or tags offered by user. Since co-mentioning in reviews emphasizes the comparison in the consumer's perspective, and that's the most important part, to fuse consumer point of view into our decision making. By Table 4.3 shows that it helps even more in listwise ranking tasks.

By comparing those three methods, we can see that the extracted and processed aspect using InstructABSA out performs other methods. We believe that the aspect group might be more down-to-earth, hence, creating more precise of parameters to describe the features.

4.5 Additional Evaluation Experiments

4.5.1 Evaluating features retrieved by review-based competitiveness mining and feature selection

In this experiment, our objective is to ascertain the relative contributions of competitiveness features, thereby determining whether it is imperative to incorporate both or if doing so introduces extra noise into our model. We separate the experiment into four parts: baseline, baseline with degree of competitors, baseline with competitiveness feature, and all added together. All experiments are conducted on the three methods promoted in RESPECT. The results shown in Table 4-4, Table 4-5 and Table 4-6:

Table 4-4 Result of evaluating features of RESPECT (LDA)

Method	**Model**	**precision@1**	**precision@3**	**MRR**	**NDCG**
Baseline (Wang et al., 2016)	LGBMRanker	44.59%	36.31%	59.23%	69.11%
RESPECT(LDA) w/o Competitiveness Features (LDA)		44.66%	36.46%	59.41%	68.99%
RESEPCT(LDA) w/o Degree of Competitors		47.05%	37.37%	61.83%	70.78%
RESPECT (LDA)		58.12%	43.94%	70.20%	76.41%

Table 4-5 Result of evaluating features of RESPECT (pre-defined)

Method	Model	precision@1	precision@3	MRR	NDCG
Baseline (Wang et al., 2016)	LGBMRanker	44.59%	36.31%	59.23%	69.11%
RESPECT(pre-defined) w/o Competitiveness Features (pre-defined)		44.66%	36.46%	59.41%	68.99%
RESEPCT(pre-defined) w/o Degree of Competitors		54.29%	37.39%	65.92%	71.43%
RESPECT (pre-defined)		50.76%	39.33%	65.03%	73.11%

Table 4-6 Result of Evaluating Features Retrieved by RESPECT (InstructABSA)

Method	Model	precision@1	precision@3	MRR	NDCG
Baseline (Wang et al., 2016)	LGBMRanker	44.59%	36.31%	59.23%	69.11%
RESPECT(InstructABSA) w/o Competitiveness Features (InstructABSA)		44.66%	36.46%	59.41%	68.99%
RESEPCT(InstructABSA) w/o Degree of Competitors		62.18%	48.17%	73.79%	80.03%
RESPECT (InstructABSA)		74.86%	53.78%	83.18%	87.21%

Table 4-7 Result of Evaluating Feature Selection on RESPECT (InstructABSA)

Method	Model	precision@1	precision@3	MRR	NDCG
RESPECT(InstructABSA)	LGBMRanker	67.32%	49.89%	75.88%	82.63%
RESPECT(InstructABSA) With Feature Selection		74.86%	53.78%	83.18%	87.21%

The experimental findings presented in these table provide further evidence that these competitive features play a crucial role. The inclusion of these features resulted in a significant enhancement in the scores, underscoring their utmost importance in our research. Also, the combination of competitiveness features and degree of competitors in some sense show synergy between these two features, because only using the degree of competitors by itself did not make much improvement, but by adding both together made a considerable progress. Finally, we compare the result before and after using feature selection, numbers show significantly improvement after feature selection. The removal of irrelevant aspects greatly enhanced the performance of our ranking model.

4.5.2 Evaluating locations with different shop density

Intuitively, restaurants located in the popular area (high restaurant density) are easier to guess. Not only the location presents agglomeration towards consumers acting like a business center luring consumers to flow in, but also contains more information for us to analyze whether the place is suitable for us to open a new branch. Shown in Figure 4-4 we can see that the restaurant count in every location is longtail distributed, indicating the imbalance distribution of restaurant in our dataset.

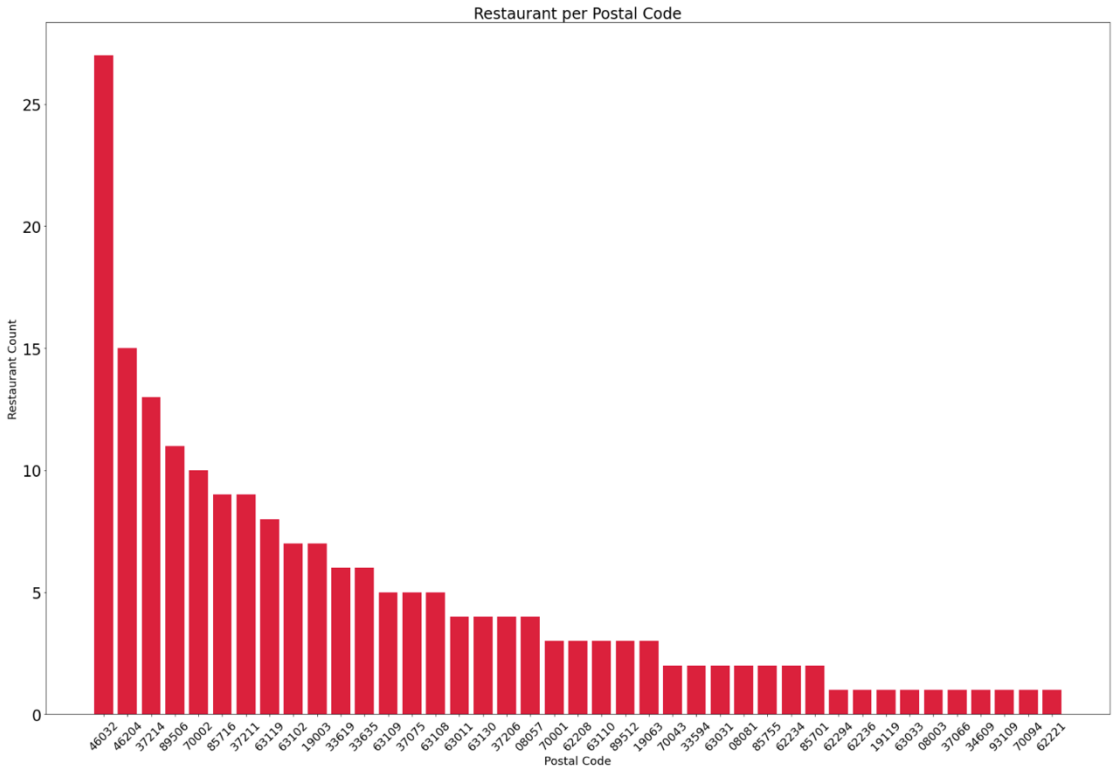


Figure 4-4 Distribution of Restaurants in Different Postal Codes

We use our best performed model – RESPECT (InstructABSA) to construct this experiment, result of this experiment is in Table 4-8:

Table 4-8 Result of Evaluating Locations with Different Shop Density

Location	Restaurant Count	Model	precision@1	precision@3	MRR	NDCG
Top 33%	305	LGBMRanker	81.07%	56.50%	89.95%	91.04%
Mid 33%	60		57.50%	39.54%	68.19%	77.85%
Bottom 33%	36		43.50%	35.67%	57.61%	70.39%

Predictions for restaurants located in popular areas show promising performance, and it is noteworthy that even with a significant disparity of over 10 times in restaurant count, the predictions remain fairly accurate. Base on this finding, we can highlight that the benefits of location-based analysis are not limited to urban or popular areas alone. The

application of location-based insights extends to rural or less populated regions as well, aiding restaurants in identifying favorable and prosperous locations. Even in remote or less frequented areas, RESPECT can provide valuable information to assist restaurants in finding optimal sites with good potential for success. This underscores the versatility and effectiveness of review-based strategies in guiding restaurant location selection, regardless of the level of urbanization or popularity of the area.

4.5.3 Evaluating Percentage of Review Used

In our dataset, the review count per restaurant also shows a long-tailed distribution, meaning that the reviews are extremely unbalanced between restaurants. This may affect the performance of our ranking model, hence, two experiments are designed to testify if review count has certain impact on our model. Figure 4-5 is used to demonstrate the distribution of review count per restaurant.

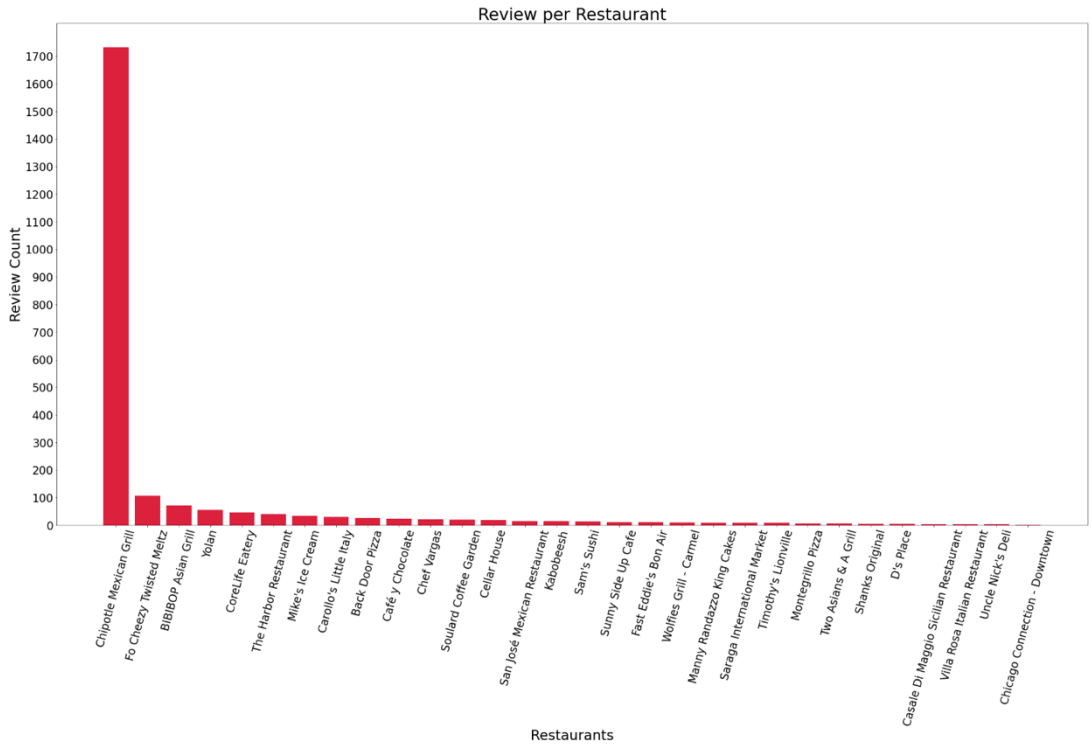


Figure 4-5 Distribution of Review in Different Restaurant

The first experiment takes a general look over the impact of numbers of review provided.

To achieve this, we employ different percentage of reviews to represent restaurants in the testing dataset. By conducting these evaluations, we aim to discern the extent of influence that the number of reviews exerts on the overall performance of our model.

Review used	Avg. Review Count	Model	precision@1	precision@3	MRR	NDCG
100%	80.69	LGBMRanker	74.86%	53.78%	83.18%	87.21%
80%	64.45		68.60%	46.59%	79.20%	85.27%
40%	32.22		64.67%	45.48%	76.55%	83.06%
20%	16.11		63.95%	45.32%	75.67%	82.79%
10%	8.06		62.20%	45.16%	75.32%	81.87%

This investigation sheds light on the importance of the quantity of reviews in shaping the effectiveness and accuracy of our predictive model, by observing the result above, we can conclude that our proposed model is insensitive towards the number of reviews, under certain amount of reviews the model is capable to accurately rank the best location.

Chapter 5 Conclusion

5.1 Summary

The key contribution of our research is novel feature provision for new store location ranking including ABSA processed features and **incorporation of community detection for identifying competitors.** This novel set of features specifically designed to aid the task of ranking potential locations for restaurants. ABSA processed features take into account various factors such as sentiment on aspect level providing valuable and detailed insights into the suitability of each location. We leverage community detection techniques to identify and analyze competitors in the restaurant industry. Unlike conventional approaches that solely rely on direct category matching, our method delves deeper into the network structure and relationships between establishments, revealing hidden patterns of competition. The results of our analysis show that this novel approach outperforms conventional methods, providing valuable insights into the competitive landscape and offering restaurant owners a more comprehensive understanding of their market rivals.

By combining these two contributions, our research delivers a comprehensive and effective framework for restaurant location selection and competitive analysis. The integration of novel features and community detection not only enhances the accuracy of location ranking but also provides valuable intelligence to guide strategic decision-making for restaurant businesses.

5.2 Future Work

For starters, dealing with word mismatch and disambiguation is worth taken into consideration. In our experiment, aspect extracted from InstructABSA is processed through heuristic basis, including choosing the aspect groups, classification criteria for

fitting words into certain groups. A systematic or rule-based approach can be taken to group these aspects. Co-mentions in review have not been lemmatize, some abbreviation of restaurant name may be left out, hence, incomplete co-mention matrix is constructed in our experiment. For example, ‘Token Ramen & Kung Fu Tea’ is mentioned as ‘Token Ramen’ and hence wasn’t count in the matrix.

Another approach might be handling multi-lingual reviews for ABSA. The review we collected from Yelp, were all in English, so if international branch were to open, how to retrieve aspects and sentiment analysis on multi-lingual basis is needed.

Cold start problem handling is also worthy of exploration. As we observe in the experiments, review count stands a critical role in the success of the ranking problem, restaurants with few reviews or few branches may not provide enough information to compute attractiveness and competitiveness feature not to mention competitor identification. Handling these niche restaurants is also a path to improve our experiment. To further expand the cold start problem, new open stores with zero information about themselves is also a compelling subject for investigation.

References

Ahedo, V., Santos, J. I., & Galán, J. M. (2021). Knowledge transfer in commercial feature extraction for the retail store location problem. *IEEE Access*, *9*, 132967-132979.

Damavandi, H., Abdolvand, N., & Karimipour, F. (2019). Utilizing location-based social network data for optimal retail store placement. *Earth Observation and Geomatics Engineering*, *3*(2), 77-91.

Fox, E. J., Postrel, S., & McLaughlin, A. (2007). The impact of retail location on retailer revenues: An empirical investigation. *Unpublished manuscript, Edwin L. Cox School of Business, Southern Methodist University, Dallas, TX*.

Ge, D., Hu, L., Jiang, B., Su, G., & Wu, X. (2019). Intelligent site selection for bricks-and-mortar stores. *Modern Supply Chain Research and Applications*, *1*(1), 88-102.

Guo, B., Li, J., Zheng, V. W., Wang, Z., & Yu, Z. (2018). Citytransfer: Transferring inter-and intra-city knowledge for chain store site recommendation based on multi-source urban data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(4), 1-23.

Hernandez, T., & Bennison, D. (2000). The art and science of retail location decisions. *International Journal of Retail & Distribution Management*, *28*(8), 357-367.

Jensen, P. (2006). Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E*, *74*(3), 035101.

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., & Mascolo, C. (2013). Geo-spotting: mining online location-based services for optimal retail store placement. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining,

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Li, C.-T., Chen, H.-Y., & Zhang, Y. (2021). On exploring feature representation learning of items to forecast their rise and fall in social media. *Journal of Intelligent Information Systems*, *56*, 409-433.

Li, Y., & Liu, L. (2012). Assessing the impact of retail location on store performance: A comparison of Wal-Mart and Kmart stores in Cincinnati. *Applied Geography*, *32*(2), 591-600.

Morrison, P. A., & Abrahamse, A. F. (1996). Applying demographic analysis to store site selection. *Population Research and Policy Review*, *15*, 479-489.

Rahman, M. K., & Nayeem, M. A. (2017). Finding suitable places for live

campaigns using location-based services. Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data,

Scaria, K., Gupta, H., Sawant, S. A., Mishra, S., & Baral, C. (2023). Instructabsa: Instruction learning for aspect based sentiment analysis. *arXiv preprint arXiv:2302.08624*.

Sokol, V., & Jordanov, K. (2020). Site selection for small retail stores using sustainable and location-driven indicators: Case study: Starbucks coffee shops in Los Angeles. In.

Taneja, S. (1999). Technology moves in. *Chain Store Age*, *75*(5), 136-138.

Wang, F., Chen, L., & Pan, W. (2016). Where to place your next restaurant? Optimal restaurant placement via leveraging user-generated reviews. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management,

Xu, M., Wang, T., Wu, Z., Zhou, J., Li, J., & Wu, H. (2016). Demand driven store site selection via multiple spatial-temporal data. Proceedings of the 24th acm sigspatial international conference on advances in geographic information systems,

Yang, H., & Li, K. (2022). PyABSA: open framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2208.01368*.

Yu, Z., Tian, M., Wang, Z., Guo, B., & Mei, T. (2016). Shop-type recommendation leveraging the data from social media and location-based services. *ACM transactions on knowledge discovery from data (TKDD)*, *11*(1), 1-21.

Zeng, J., & Tang, B. (2019). Mining heterogeneous urban data for retail store placement. Proceedings of the ACM Turing Celebration Conference-China,

Zhdanov, P., Glukhov, G., & Derevitskii, I. (2022). Machine Learning Approach to Optimal Restaurant Location Search. 2022 IEEE 16th International Conference on Semantic Computing (ICSC),