Restaurant Recommendation System using Natural Language Processing(NLP)

Soe Zhao Hong

SESSION 2021/2022

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
APRIL 2022

Restaurant Recommendation System using Natural Language Processing(NLP)

BY

Soe Zhao Hong

SESSION
2021/2022

THIS PROJECT REPORT IS
PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT
FOR

BACHELOR OF COMPUTER SCIENCE
B.CS (HONS) DATA SCIENCE

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY

APIRL 2022

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

_____Zhao Hong_____

Name of candidate: Soe Zhao Hong
Faculty of Computing & Informatics
Multimedia University

Date: 22: 04: 2022

Acknowledgment

First of all, I would like to express my most profound sense of appreciation to my supervisor, Dr. Goh Hui Ngo. She continuously contributes her time as a researcher and more critically a capable teacher. I truly acknowledged that Dr. Goh is continuously there for me to lead me within the redress heading in spite of the fact that she is exceptionally active with her obligations. Her direction and information sharing, which have given me the capability to total the work.

Other than that, I would like to appreciate my course mates who continuously share their experiences and advice with me, helping me to solve my issues and questions. Too, I am thankful to Multimedia University College (MMU) which gives me a fabulous learning environment to memorize and prepare my abilities.

Lastly, I would not have had the chance to study at MMU to gain professional knowledge without my parents who always supported and encouraged me, this is critically important which helps me through the difficulties. So, I have a profound appreciation for them.

Abstract

Recommendation system has been implemented in applications nowadays especially E-commerce applications since recommending preferable items among a huge number of items is essential on the online market. To recommend preferable items to users, knowing users' opinions is very important instead of rating. In this paper, we have proposed a content-based restaurant recommendation system using Natural Language Processing (NLP). Words embedding method, a subfield of NLP: TF-IDF, BERT, CountVectorizer, and Word2Vec is used to represent the users' reviews, followed by finding the similarity between restaurants using cosine similarity. In such a list restaurants will be produced based on the cosine similarity score from the restaurants. As for the evaluation, this work employs the supervised proxy problem, in such, BERT has outperformed the other proposed methods in the evaluation of 3 classification models: Random Forest, Naive Bayes, and KNN. Besides, the food entity from the review is found using a pre-trained model from DeepNote.com and uses some specific Part of Speech to define the food entity's sentiment. Last, the result in phase 1 and phase 2 will be combined in order to provide a complete recommendation system.

Table Content

List of Tables

List of Figures

Chapter 1 Introduction

In the world of multimedia and the growing amount of electronic services, choosing a suitable product has become complicated and takes a lot of time. If a user faces a bunch of products or services, the user feels very bored and weary to choose a preferred product. So, the eCommerce applications such as Grab, Foodpanda, MyFamily Mart, etc have proposed solutions to solve the problem. For instance, advertisements, promotions, and recommendations are provided in these applications. The recommendation system improves the user experience as it provides a product that meets the preferences of a user based on the historical transactions or the preferences of other users which might be the same as the user. The recommendation system also can save time and money used by users on choosing suitable and interesting products. Therefore, the recommendation system has become more important to provide a better experience for people's internet life.

The covid-19 virus sweeps through the whole world in the early of 2020 and it takes many lives of humans in a very short time. This made a lot of countries start to alert and make many strategic decisions to confront the Covid-19 virus. Due to the ways of the diseases spreading, a new norm has happened in the human lifestyle. During the pandemic period, everyone puts on a mask and keeps their distance from each other to prevent contraction. Besides, the operating mode of all companies is changed also to avoid contact between workers. For example, the workers are not allowed to work in the office but work from home. Then, this has made changes to people's lifestyles. Human activities are restricted like outdoor activities are not allowed, cannot dine in restaurants but only can take away or ask for delivery services from the restaurant, and social events are banned. Due to this situation, online applications have become more and more important in people's daily life such as social media, shopping applications, food ordering applications, etc. During this period, these applications become a necessary application for a human to handle daily life especially the food ordering application. For example, Grab Food, Food Panda, EASI MY, Uber Eats, etc. To meet the interest of the user while using these food order applications, the recommendation system becomes a very important role in the applications to find out the best combination of the food that the user likes through the

users' personal preferences. In addition, these online applications always collect the rating and reviews from the users after enjoying the products. It is because the rating of products from users presents the satisfaction of users on the products that they ordered but the reviews will totally present the manner and emotion of users to the products and sellers through the typing words. So, reviews can present the opinion of users more complicated than the rating. Then, it can be concluded that the recommendation based on the products' reviews of users can provide a more humanized recommendation to the users.

Currently, the recommendation methods have been categorized into three basic types: content-based filtering recommendation (CB) (J Ananda Babu et al., 2021), collaborative filtering recommendation (CF) (J Ananda Babu et al., 2021), and hybrid recommendation (Bagher et al, 2017). The CB recommendation is a recommendation system based on the content information of the products or user profiles, but the CF recommendation recommends based on two types of patterns which are user-based, and item-based. The user-based CF recommends products to a user which similar users have liked, but item-based CF recommends similar products based on how people rated them in the past. CF cannot provide better performance while it does not have enough users' information (Luong Vuong Nguyen, 2020) because it cannot recommend products without any previous user's data. So, the limitation of CF is cold-start, sparsity, and scalability problems. Then, the overspecialization and limited content analysis are the limitations of CB. Last, the hybrid recommendation is the combination of two or more patterns of recommendation methods like a combination of CF and CB, etc. The hybrid recommendation will take the advantage of both recommendation methods and cover the limitations of both methods. Through these recommendation methods, each has its advantages and limitations.

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech in the same way a human being can (Gobinda G. Chowdhury, 2003). NLP is a combination of computational linguistics, statistical, machine learning, and deep learning models to help computers on process and understand human language in the form of text or voice data, complete with the speaker or writer's intent and sentiment. NLP can drive computer programs that change

text from one language to another by responding to spoken commands, and summarize a huge volume of text rapidly. As such, if a recommendation based on reviews is implemented into a recommendation system, NLP will play an important role in understanding the reviews of users.

1.1 Problem Statement

In this study, we plan to build a restaurant recommendation system based on the customers' reviews of particular restaurants. The reviews are rich and provide a lot of information such as ambiances, favor of the food and beverage, etc. In the hope that customers' needs can be found in the reviews, not just based on the numerical rating.

Reviews are the comments about the restaurant, so a content-based recommendation will be built. To process the reviews data, Word Embedding methods such as Term Frequency - Inverse Document Frequency (TF-IDF) (Tessy Badriyah et al., 2018), Bidirectional Encoder Representations from Transformers (BERT) (Xusong Chen et al., 2019), CountVectorizer, and Word2Vec (Hao Chang et al., 2021) which is a sub-field of the Natural Language Processing (NLP) will be used to recommend similar restaurants to the user, followed by calculating the similarity between restaurants, the Cosine Similarity matrix (Joanna Cristy Patty et al., 2018) will be used and the model will be employed the supervised proxy problem by training classification models: Random Forest Classification (Ayush Singhal et al., 2017), Naive Bayes (S.M. Asiful Huda et al., 2019), and K nearest neighbor (KNN) (Ayush Singhal et al., 2017) to evaluate the recommendation system. The classification models will be compared to find the best performance word embedding methods by following the Precision, Recall, Accuracy, and F1-score (Maria del Carmen Rodriguez-Hernandex et al., 2020) from the classification models.

After that, we will focus on identifying and extracting food named-entity and its sentiment in each review. As for a complete recommendation, sentiment analysis (S.M. Asiful Huda et al., 2019) and named entity recognition will be used in phase 2 to recognize the food name in the review and define the food taste from the customers' reviews. Then, the trained model in phase 1 will be integrated with the phrase 2 models in order to provide a complete recommendation system.

1.2 Objectives

1. To propose an algorithmic framework for a restaurant recommendation system
2. To employ Natural Language Processing (NLP) for the recommendation of restaurants.
3. To perform evaluations on the models to validate the performances of proposed approaches.

1.3 Project Scope

During phase 1, we aim to build a content-based recommendation using a few word embedding models, followed by calculating similarity using cosine similarity. The dataset used is not a benchmark dataset, hence, performance evaluation is carried based on the word embedding models.

In phase 2, our aim is to recognize the food name and its sentiment from the review. Named entity recognition and sentiment analysis will be carried out. Then, a complete recommendation system will be produced, integration of phrase 1 and phase 2.

1.4 Timeline

Phase 1

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Literature Review						
Analyze dataset						
Survey						
Build framework						

	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Evaluation						
Writing report						

Phase 2

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Enhance framework						
Implement GUI						

	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Implement GUI						
Testing						
Finalize report						

Chapter 2.0 Literature Review

2.1 Existing Online System

Many applications nowadays are trying to build a recommendation system to meet the preferences of the users, for instance, social media platforms, online stores, and food ordering systems, etc. In these applications, a recommendation system is used to promote products or advertise based on the location and interest of the users. Due to such a recommendation system, it makes people's lives convenient as the users can easily find out the necessity or interesting product that the user wants in a short time.

2.1.1 Grab (Mobile Application / Website)

Grab is an application that contains a lot of services via a mobile app. Grab started in 2012 as the MyTeksi app in Malaysia and expanded into other services after changing the application name to GrabTaxi. In this application, there are various services for the users. For example, delivery service, transportation services, insurances, etc. Based on Figure 2.1 below, some recommendation systems are shown.

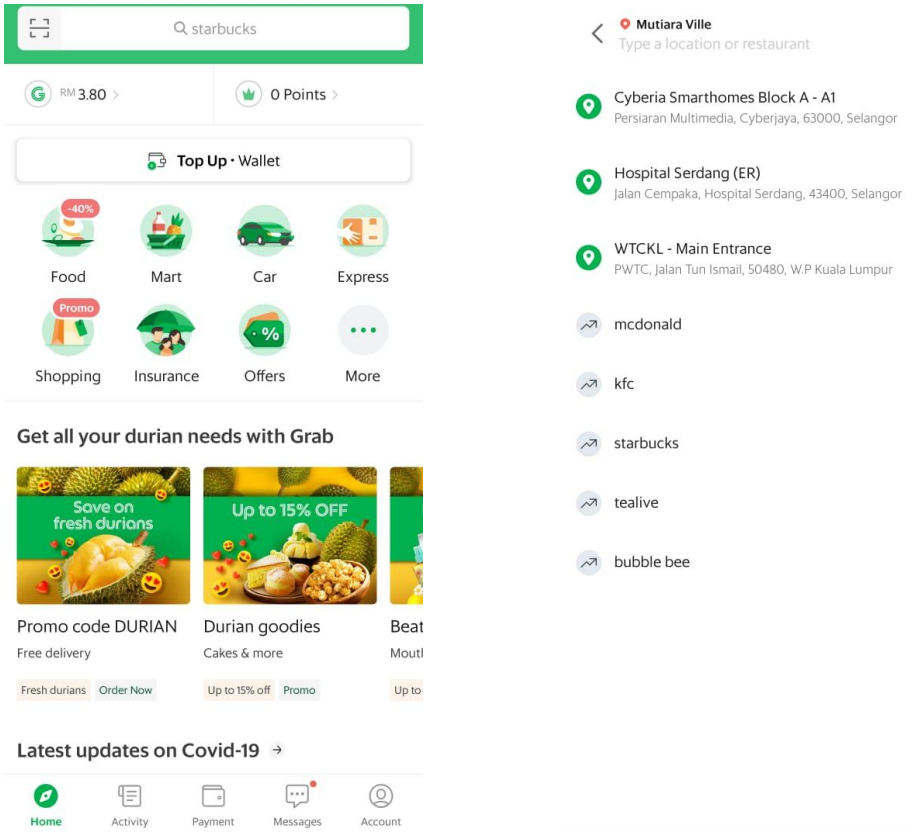

Figure 2.1 (a) & (b): Popular trends are always searched by users.

In Figure 2.1 (a), there is a shop named 'Starbucks' that has been recommended to the user in the search bar which is on the top of the application. Besides, some places also recommended by the system are Mcdonald's, KFC, Starbucks, Tealive, and Bubble Bee in Figure 2.1 (b). Based on the observation, these places are always searched by the Grab users in this particular location which is provided by the user.

2.1.1.1 Grab Food

Grab Food is a meal delivery company that operates as a subsidiary of Grab which was established in 2015. The Grab Food service is offered through GrabApp and provides services in several countries, including Singapore, Indonesia, Vietnam, and Malaysia. This application provides two types of version which is mobile application and website. Based on Figure 2.2 (a), (b), and (c), some recommendation is provided by the application to the user.

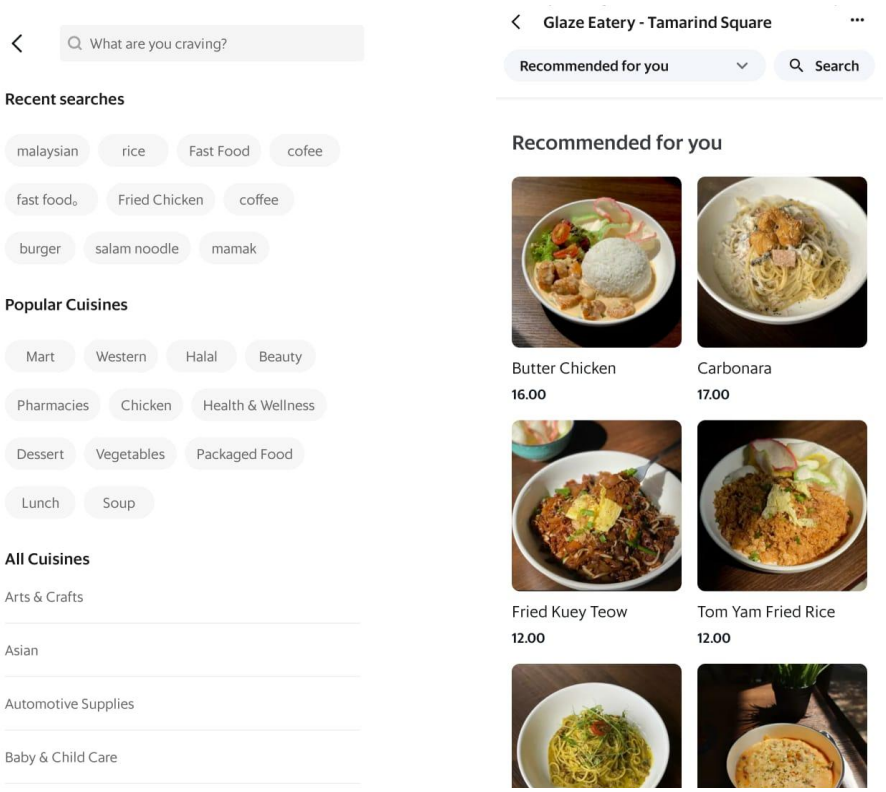

Figure 2.2 (a) & (b): The mobile application of GrabFood shows the popular cuisines and popular food.

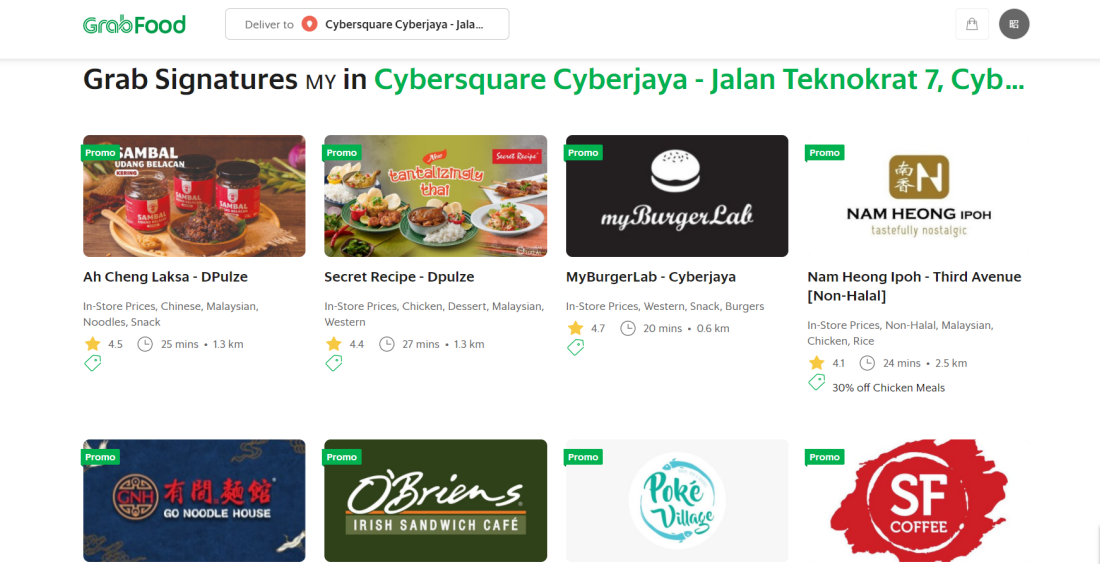

Figure 2.2 (c) : The website of GrabFood shows some promotions at the location of the users.

Figure 2.2 (a) shows the searching function of the application, some tags are shown below the category 'Popular cuisines' which is the cuisines that are always searched by the users in this application. Besides, Figure 2.2 (b) presents the popular food that is frequently ordered by the customers within the same restaurant. In Figure 2.2 (b), Some promotions provided by the restaurant will be shown to users in the specific position of the main menu on the website based on the location of the user.

2.1.2 My FamilyMart Online (Mobile Application)

My FamilyMart Online is an eCommerce application that was released by FamilyMart in November 2020. This application is built to let FamilyMart's customers enjoy the products from this company without going out due to the pandemic situation.

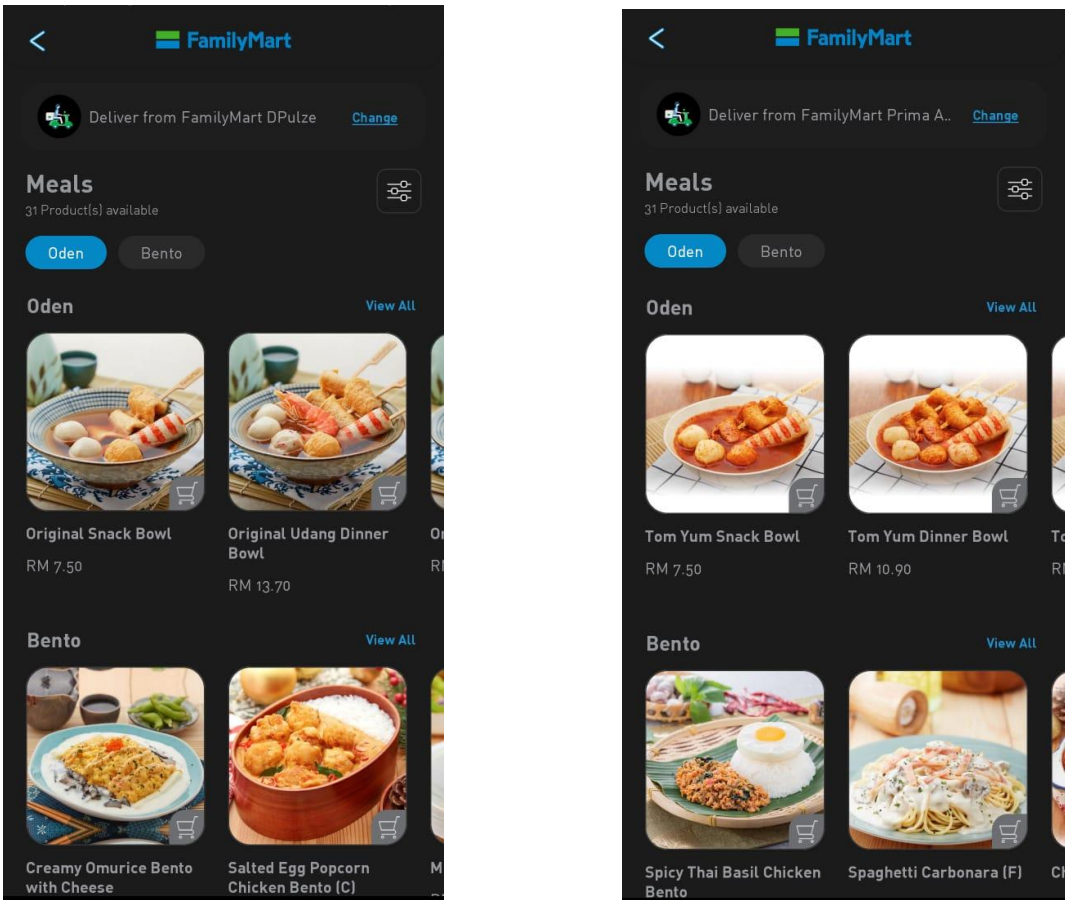

Figure 2.3 (a) & (b): Different markets present products based on users' preferences within the specific area.

In Figure 2.3 (a) and Figure 2.3 (b), the application presents the products in the specific category are different. Based on the observation, the systems will recommend the products based on the users' preferences in the particular area. So that the new customer can view the products that are popular within this area while the first time visiting the market and the old customers can find out favorite products conveniently. Based on Figure 2.3 (c), The trend that is always searched by the users in this application is also recommended to the user while using the search function in the My Family Mart Online application.

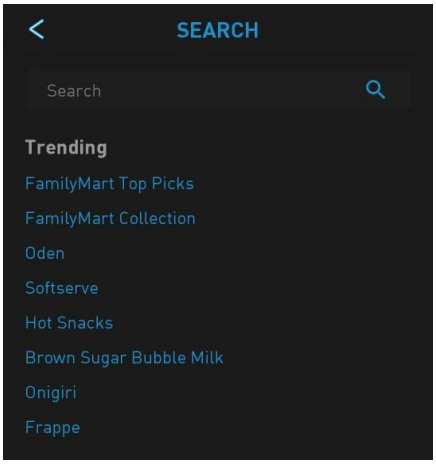

Figure 2.3 (c): The search function shows the trend that is always searched by the user in the application.

2.1.3 Foodpanda (Mobile Application / Website)

Foodpanda is an online food and grocery delivery platform and a subsidiary of Delivery Hero. This platform was built in November 2013 and currently is the largest food and grocery delivery platform in Asia. In Figure 2.4 (a), some restaurants are recommended to the user in the "Panda Picks" category which is usually ordered by the user at that particular location. Besides, the cuisines that are popular in the restaurant will be presented on the main menu of the restaurant based on Figure 2.4 (b).

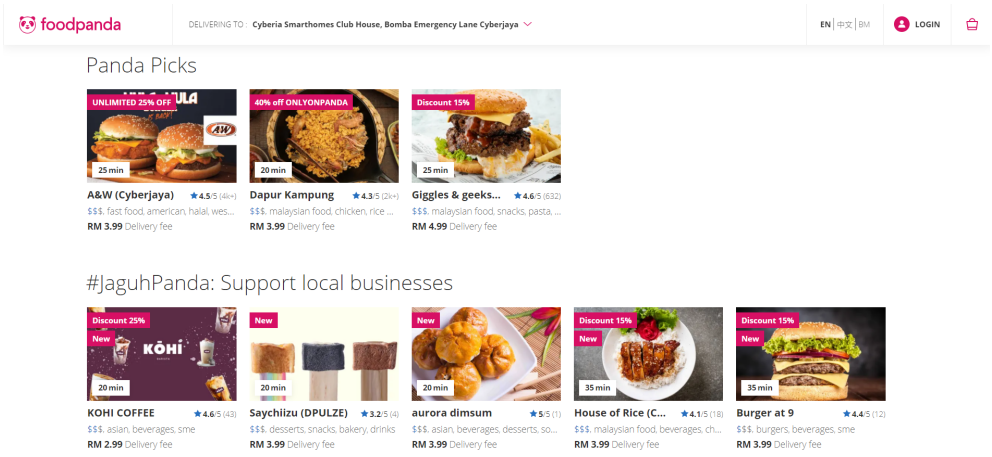

Figure 2.4 (a): Recommendation of the restaurant in the main menu of the Foodpanda.

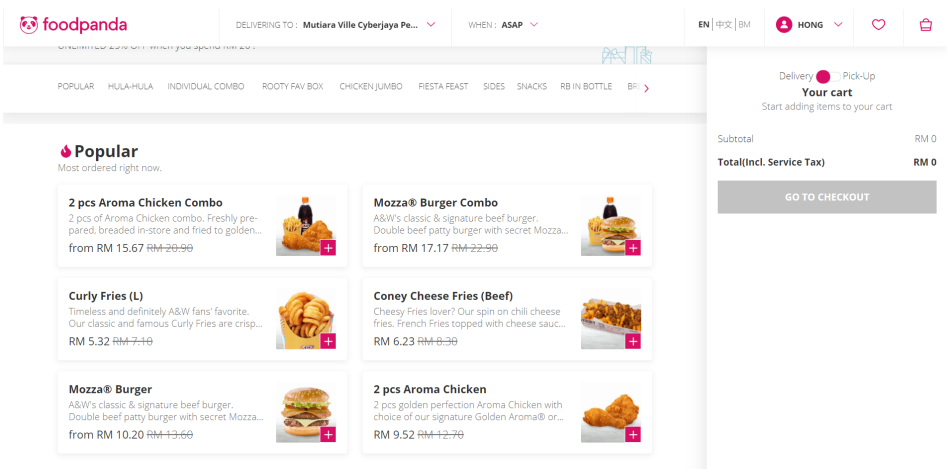

Figure 2.4 (b): Food that is usually ordered by customers will be set as popular on the restaurant menu page.

Same with other applications that contain a recommendation system, the FoodPanda application also has recommended the tags that are always searched by other users in this application which is shown in Figure 2.4 (c). Besides, Market Basket Analysis is also used in this application based on the observation in Figure 2.4 (d). Through the combination of the meals and drinks from the users in this application, the application list out a few drinks that are always ordered with users' meals.

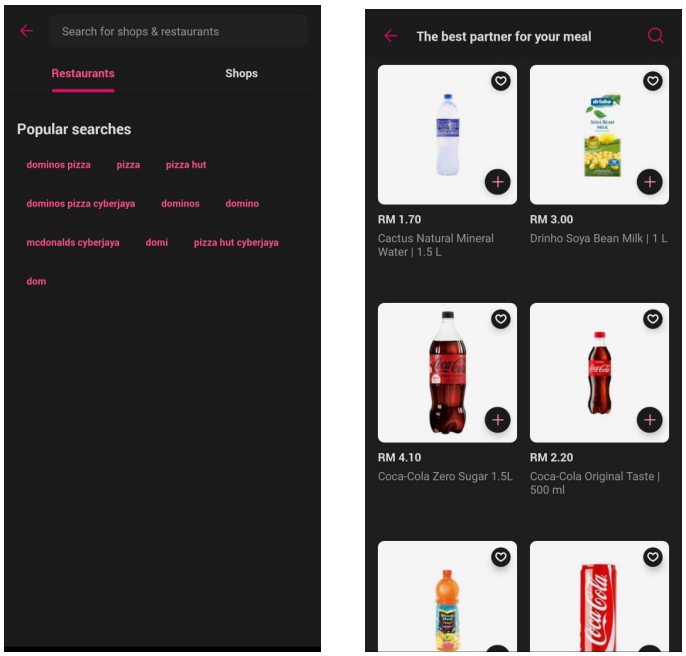

Figure 2.4 (c): The searching function recommends the tags that are always searched by the users.Figure 2.4 (d). The best partner for the users' meal is recommended.

2.1.4 EASI MY (Mobile Application)

EASI MY is a food delivery platform only available in Malaysia which is released in May 2018. In this application, there is a category “Picked for you”, the recommendation page that is based on user location.

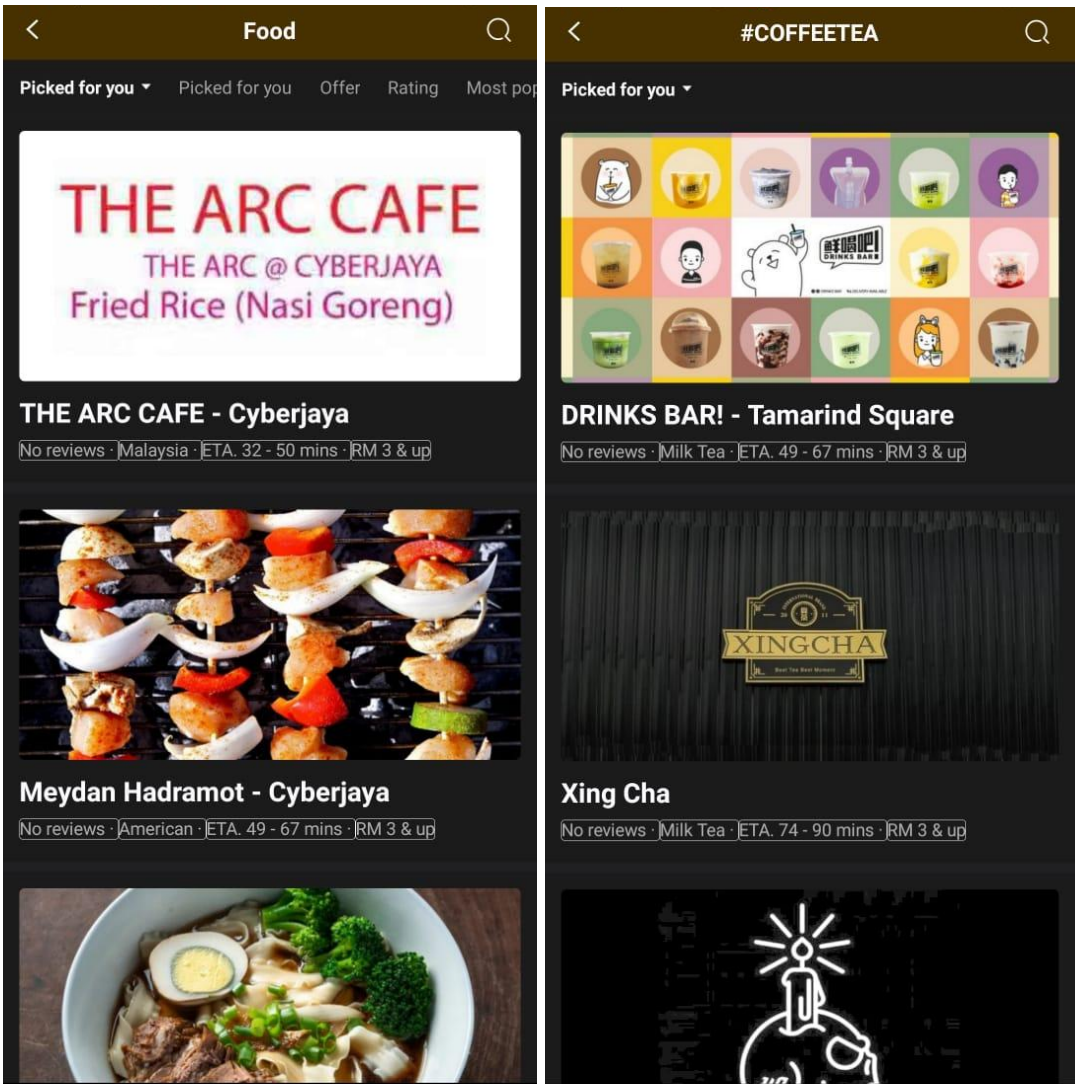

Figure 2.5 (a) & (b). Restaurant page that is recommended to the user.

2.1.5 Pop Meals (Mobile Application & Website)

Pop Meals is formerly known as DahMakan and Dahmakan was rebranded in late 2020. This application is a startup company that offers ready-to-eat meals for delivery and provides a variety of cuisines, drinks, and sides from international. In Figure 2.6(a), the main menu of the Pop meals website recommends the new cuisines and best sellers on this platform. Therefore, the application of Pop meals provides a ranking of the food which is the most popular food among the users based on Figure 2.6 (b).

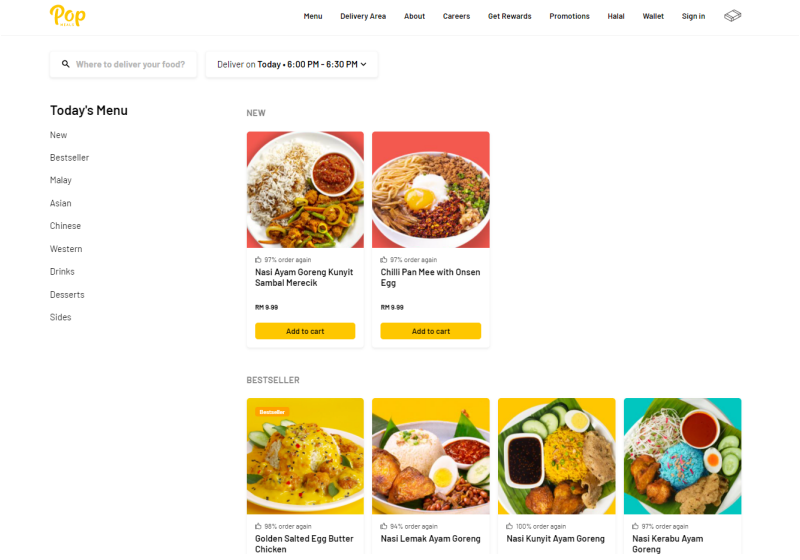

Figure 2.6 (a): Main Menu of Pop Meals Website that shows the new food and the food that is the best seller.

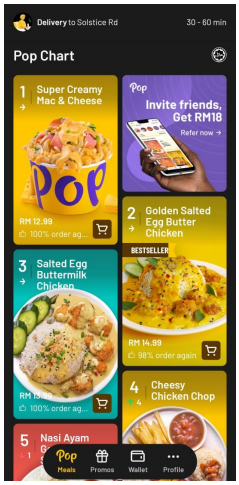

Figure 2.6 (b): Main Menu of Pop Meals Application that a ranking of the food.

2.1.6 Shopee (Mobile Application/ Website)

Shopee is an online shopping platform from a company named Shopee Pte Ltd in Singapore since June 2015. This platform currently serves consumers and sellers throughout Southeast and East Asia and several countries in Latin America.

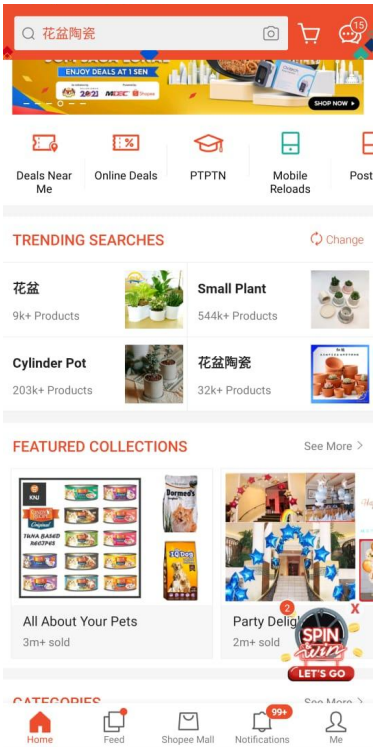

Figure 2.7 (a): The recommended tag is shown in the search bar to the user and recommends products based on the history search.

Based on Figure 2.7(a), there is a tag shown in the search bar and the tag is related to the history search of the user. For example, ‘cactus’ is searched by the user before, and ‘small clay pot’ will be shown in the search bar. A category ‘Trending searches’ will be recommended to the user based on the history search of the user. In this category, some random products which are related to the history search of the user will be recommended to the user. Besides, there is another category named ‘Top Products’ that shows the products which are similar to the history bought, but these are the top sales products in the category of the history search (Figure 2.7(b)).

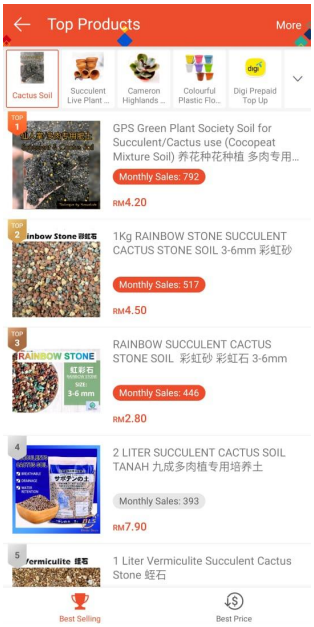

Figure 2.7 (b): The top products are similar to the product that was bought by users before.

In the shopping cart of the Shopee, some random products will be recommended to the user which is related to the history bought by the user. For example, a cactus plant is bought by the user and a few types of soil will be recommended to the user. On a particular product page, a few similar products that are related to that product will be recommended to the user. Examples are shown in Figure 2.7 (c) & (d).

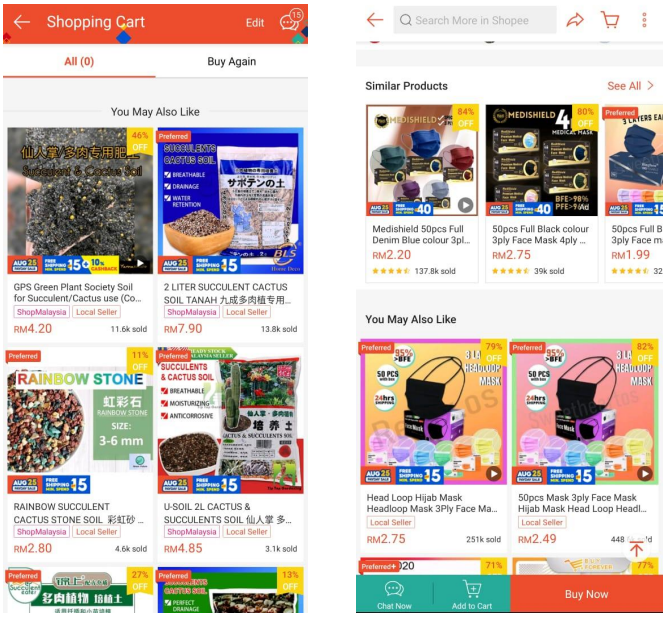

Figure 2.7 (c): Recommends product based on the history bought by the user. Figure 2.7 (d). The recommended product is related to the particular product.

2.1.7 Lazada (Mobile Application/ Website)

Lazada is an Online Shopping Application that is owned by Alibaba Group and was built in June 2013. Countries that support this application are Indonesia, Malaysia, Philippine, Singapore, Thailand, and Vietnam. This platform was claimed as the top e-commerce platform in Southeast Asia in 2019.

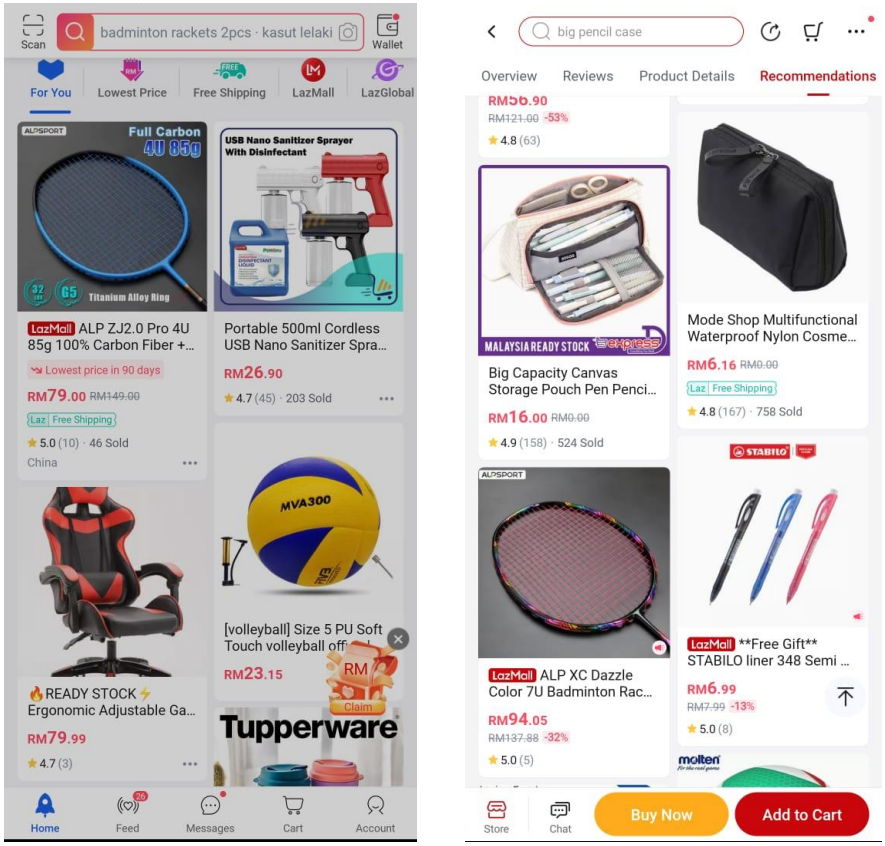

Figure 2.8 (a): Recommended products that relate to the history search of the user. Figure 2.8(b). Recommend products that relate to the particular product.

Based on the observations from Figure 2.8 (a), Lazada will recommend products based on the search history of the user or the trending search of users Lazada. For example, the user searched for sports equipment and gaming chairs before. Then, Lazada recommends a few sports equipment, chairs, and sanitizer to the user based on Figure 2.8 (a). The sanitizer is recommended because of the pandemic Covid-19 now and this type of product will always be searched by other users. Besides, some similar products are also recommended on a particular product page. For example, a pencil case is searched and stationery is recommended.

2.1.8 Facebook (Mobile Application/ Website)

Facebook is a social media platform and it was built in February 2004. This social media platform is very famous and popular with people nowadays. All information from all over the world can be received from this platform. This is the reason why Facebook has accumulated about 2.85 billion users.

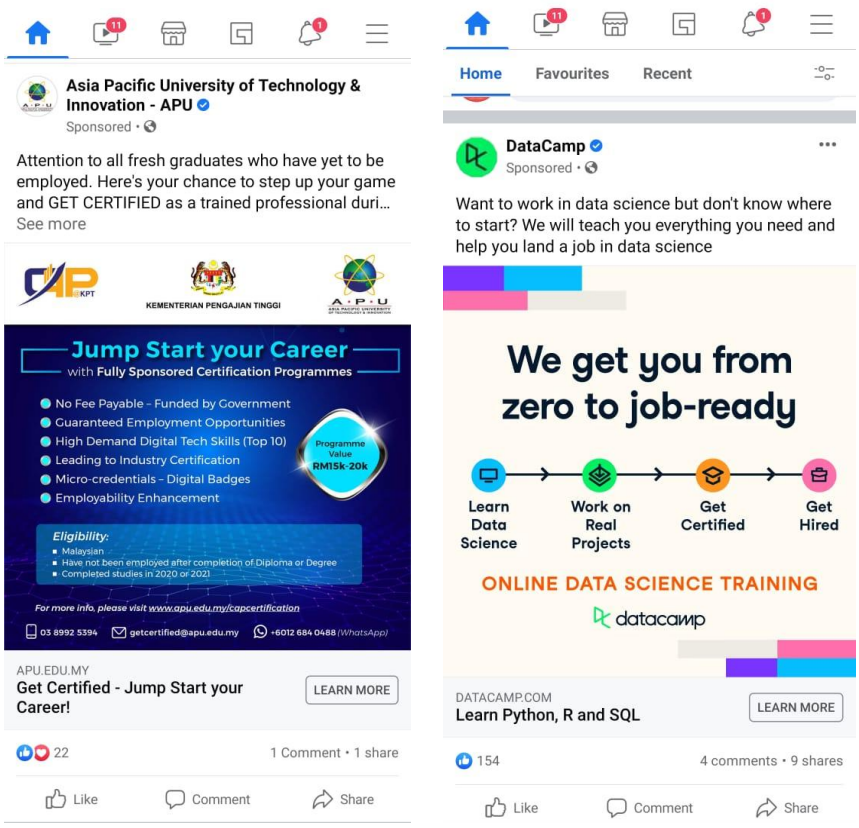

Figure 2.9 (a) & (b): This shows the advertisement that is related to the user's browser history.

Google and Facebook have agreed to team up to work together in online advertising since 2020. Based on the observations in Figure 2.9 (a) and (b), the user has been recommended a Computer Science program because the user searches for some knowledge about Computer Science in Google recently.

2.1.9 Instagram (Mobile Application)

Instagram is also a very popular social media platform to people nowadays. This platform was released in 2012 and also has accumulated about 180 million active users. Instagram allows users to build an online shop to sell products and these shops show a blue tick behind the shops' name which indicates that the shop has been verified by Instagram.

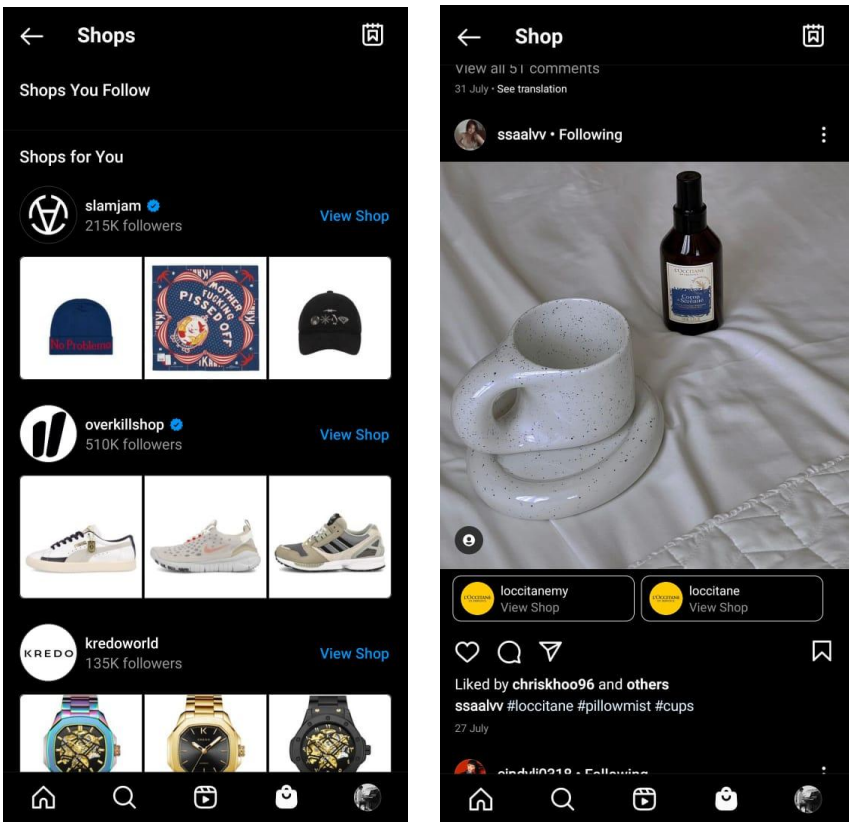

Figure 2.10 (a) & (b): Recommended shops and products that are related to the user's information.

Instagram provides a function named 'Follow' and this function can help users to keep track of the information from the user that has been 'Followed'. Based on this function, Instagram can realize the preference of the users and recommend the related products to the user (Figure 2.10 (a)). Besides, the product photo posted by the user which contains the shop tags will also be recommended to the user in the Instagram shop (Figure 2.10 (b)).

Table 1.0: Online existing system

Name	**Types**	**Year**	**Recommendation model**	**Way to recommend**
Facebook	Mobile Application / Website	2004	Content-based	Recommend content that is similar to the google search history.
Instagram	Mobile Application	2012	Collaborative filtering	Recommend products based on the preference of the user.
Foodpanda	Mobile Application / Website	2013	Location-based, Content-based	Recommend restaurants based on the user's current location. Users' search history that is popular is set as a tag to recommend in the search bar.
Lazada	Mobile Application / Website	2013	Content-based, Collaborative based	Recommend products based on the user's history search. Related products will be recommended on the particular product page.
Grab/ Grab Food	Mobile Application / Website	2015	Location-based, Content-based	Recommend a restaurant based on the location of the user. Search trends from users will recommend through the searching function.
Pop Meals	Mobile Application / Website	2015	Content-based	Recommend food based on food ranking. Food that is new to the platform will be recommended as a priority.
Shopee	Mobile Application / Website	2015	Collaborative filtering, Content-based	Recommend similar and related products based on the history search of the user.
EASI MY	Mobile Application	2018	Location-based	Recommend restaurants based on the location of the user.
My FamilyMart Online	Mobile Application	2020	Location-based, Content-based	Recommend products based on customers' preferences in that particular location. Search trends from users will also be recommended in the searching function.

2.2 Related Work

In this section, the author had done some literature reviews which are about sentiment analysis, recommendation model, and word embedding model.

2.2.2 Recommendation System

J Ananda Babu et al. (2021) had been proposed two types of recommendation systems: the content of Itering and the collective Itering. To build these recommendation systems, Term Frequency- Inverse Document Frequency (TF-IDF and doc2vec are used for measuring the film data from the MovieLens dataset, and Neighbours nearest K (KNN) is used to measure the similarity score of the film data. During the evaluation, the author found that the accuracy of the collective Itering is higher than content-based Itering and the content-based iteration is costly rather than collective Itering. However, a limitation is found which is KNN's over ratting issues and the author decides to fix it by using regularisation in the future.

Rahimpour Cami Bagher et al. (2017) propose a content-based recommendation system based on the user trend. The trend-based user model is constructed by combining user profiles into an extension of Distance Dependent Chines Restaurant Process (dd-CRP), a Bayesian Nonparametric model. The result of this work presents that the user-trend modeling for content-based recommender has excellent accuracy and improved its effectiveness over time.

Xingjuan Cai et al. (2020) built a hybrid recommendation system with a many-objective evolutionary algorithm (MaOEA) because the author thinks that the recommendation system is not only focusing on accuracy while users have different requirements. In this paper, the author has used three types of recommendation algorithms: user-based collaborative filtering, item-based collaborative filtering, and content-based & collaborative filtering (hybrid) to compare with the proposed method (RVEA). The proposed method is performing well in MaOEA which has high novelty and coverage on the basis of accuracy and diversity.

Wei-Ta Chu et al. (2017) investigate the influence of visual information on restaurants' recommendation systems by integrating 2 recommendations into a hybrid recommendation: content-based recommendation and collaborative filtering recommendation. A convolutional neural network (CNN) is used to extract out the visual information. Three recommendation methods: Factorization machines (FM), Matrix factorization (MF), and Bayesian personalized ranking matrix factorization (BPRMF) used as the baseline approaches in this work. Then, these three methods will be enchanted by adding visual information. At the evaluation part, the performance of FM and MF are increased a lot but BPRMF only increases slightly after being enchanted with the visual information. Although the performance of BPRMF only slightly increases after enchanted, BPRMF still is the best performance between these three methods. In this work, the sparsity and cold start problem of the recommendation system is solved when adding the visual information.

Monali Gandhi et al. (2019) have proposed an enhanced approach for the tourism recommendation system. The recommendation is combined with association rules mining and the collaborative filtering recommendation to improve the quality and strong recommendations to the users. The association rules mining helps to find out the best combination of output for the users and this makes the collaborative filtering recommendation system that combined with it gives out a very high performance compared with a content-based recommendation. Both methods have their own disadvantages but they can be made up for each other while both methods are combined together. So, the author decides to design a more powerful hybrid model by using more combinations of the model in the future.

Khushbu Jalan et al. (2017) built a context-aware hotel recommendation system based on a hybrid approach. This approach is proposed to solve the cold start of the recommendation model. In this paper, the author has used sentiment analysis to find out the review attributes which are positive or negative. Then, the collaborative filtering recommendation system that combines with the sentiment analysis can provide a better recommendation by preventing recommendation the output that is full of negative reviews.

Nanthaphat Koetphrom et al. (2018) proposed a work that compares 3 types of recommendation systems: content-based recommendation, collaborative filtering recommendation system, and hybrid recommendation system. The hybrid recommendation is the combination of content-based recommendation and collaborative filtering recommendation. In this paper, the content-based recommendation uses regression to create a prediction model from characteristics of customer and restaurant, but collaborative filtering recommendation employs a combination of clustering, similarity, and weight sum to investigate the factor that affects the rating from users. The result of this work is evaluated with residuals which are used to investigate the difference between the prediction and actual value and mean absolute error (MAE) is used to calculate the absolute error of prediction. Based on these evaluation metrics, the hybrid approach has performed slightly better than the other two recommendations.

Bhumika Narwani et al. (2020) propose a hybrid recommendation system by combining the content-based recommendation and collaborative filtering recommendation. Besides, geospatial location and historical check-in of users are also inserted into the recommendation system to increase the accuracy of the system. TF-IDF will be used in the system to proceed with the word embedding and Least Square Error(LSE) will be used as the predicted model to avoid model over-fitting. In this project, the content-based recommendation will be started while the preference of the users is collected in the first two weeks. The author had concluded that this proposed recommendation had improved the user experience. The geospatial location and historical check-in of users can overcome the issue of content-based recommendations: excessive specialization. Besides, the issue of collaborative filtering recommendation: cold start problem can be also overcome by making the user fill up some information while registering.

N. Pradeep et al. (2020) have proposed a content-based movie recommendation system that focuses on the cast, keywords, crew, and genres of the movies. The author has compared multiple recommendation models: collaborative filtering, content-based, and hybrid approach. The content-based recommendation system has been chosen because Content-based recommendation does not consider other user-profiles and it will make sure the user gets personalized suggestions for their input. To train this recommendation system, the author has

used TF-IDF as the word embedding method and Cosine similarity to calculate the similarity of the vector from TD-IDF.

Tomasz Rutkowski et al. (2018) present a content-based recommendation system based on a neuro-fuzzy approach. The neuro-fuzzy approach in this study is used to decide the recommendation to the user. This approach is possible to learn and simulate users' decisions based on users' actions. In the evaluation part, neuro-fuzzy is divided into two which are neuro-fuzzy pre-train and post-train to compare the effectiveness of this approach. Besides, Deep Neural Network (DNN) also compares this approach during the evaluation. Unfortunately, DNN has very high accuracy and effectiveness compared with neuro-fuzzy, but neuro-fuzzy only performs slightly worse than DNN. The author decides to combine both methods to build a high-quality recommender in the future.

Ayush Singhal et al. (2017) have made a review in the field of recommendation systems using deep learning technology. This review is made up of three parts: Collaborative Filtering system, Content-based system, and Hybrid system. The author found that the most efforts of deep learning are towards enhancing collaborative filtering. Besides, most of the deep learning development is focused on movie and music recommendations.

Table 2.1: The summary of the recommendation system literature review.

Author	**Year**	**Method**	**Dataset**	**Conclusion**
Wei-Ta Chu et al.	2017	MF, FM, BPRMF, CN, CNN	Own dataset	The results show performance superior to the state of the art.
Ayush Singhal et al.	2017	Collaborative, Content-based, Hybrid system	-	Most deep learning efforts have been towards enhancing collaborative filtering approaches.
Rahimpour Cami Bagher et al.	2017	Content-based recommender	Tweets of New York Times, BBC, and Associated Press.	The results show the accuracy of the proposed approach and its ability to evolve over time.

Khushbu Jalan et al.	2017	Collaborative filtering, Sentiment Analysis	Dataset from Tripadvisor.com	The sentiment analysis solves the cold start of collaborative filtering successfully by differences the reviews from the users are good or bad.
Tomasz Rutkowski et al.	2018	Content-Based Recommendation	MovieLens 20M Dataset, TMDB API service	Deep neural networks made a very high accuracy of the recommender compared with neuro-fuzzy performs slightly worse.
Nanthaphat Koetphrom et al.	2018	Content-based, Collaborative Filtering, Hybrid filtering	Own collected dataset	Hybrid filtering performs slightly better than content-based and collaborative recommendations.
Monali Gandhi	2019	Content-based, Collaborative filtering	Own collected dataset	The recommendation is combined with association rules mining and the collaborative filtering recommendation has high performance compared with a content-based recommendation.
Bhumika Narwani et al.	2020	Hybrid, Collaboration Filtering, Content-based, Location-based	Yelp Dataset	Collaborative filtering recommends is used during the first two weeks of the system and let the system collect preferences of the user for the content-based recommendation. To solve cold start, the location-based recommendation technique is used.
N.Pradeep et al.	2020	Content-based	Own collected dataset	The content-based recommendation does not consider user-profiles and makes sure the user gets personalized suggestions for their input.
Xingjuan Cai et al.	2020	Item-based & User-based collaborative filtering, Content-based	MovieLens 1 M Dataset	Many-objective evolutionary algorithms (MaOEA) are performing well with RVEA.
J Ananda Babu et al.	2021	Collaborative Itering, Content-based Itering	Movie Lens dataset on Kaggle	Collaborative Itering is more accurate than content-based Itering and both methods strengthen the recommendation framework.

2.2.3 Word embedding

Tessy Badriyah et al (2018) have proposed a recommendation system for property search using a content-based filtering method. In this paper, the author has collected the user behavior by collecting the searching advertising content that was searched by the user before. Then, the preferences of the user will be embedded by the TF-IDF and generate the most frequent words in the advertisement. Apriori algorithm which is one of the association mining algorithms is used to find similar combinations and suggest to the user.

Hao Chang et al. (2021) proposed a content recommendation system for homestay to overcome the problems of the application of recommendation systems in the homestay industry which cannot provide better choices for the user. In this proposed system, the TF-IDF model is trained to vectorize the features of the homestay and the cosine similarity is used to get the similarity between the homestay. The result of this work is that user satisfaction has been improved where user preferences can be met with the features of the homestay.

Xusong Chen et al. (2019) propose a BERT for session-based recommendation (BERT4SessRec) in a challenge: Content-Based Video Relevance Prediction. In this work, BERT4SessRec is trained with all session data during the pre-training stage and this proposed method has been achieved to help with the enhancement of BERT. The author decides to study the BERT by combining the content and embedding features in the future.

Joanna Cristy Patty et al. (2018) built a recommendation system for the selection of cosmetic purchases and aims to guide customers to make better decisions on suitable products. So, a content-based filtering method is used in this recommendation system. TF-IDF is used to calculate the vector of the product information and the similarity between the vectors is calculated using the Cosine similarity algorithm. This proposed method can recommend accurately to the user in the result of this paper.

Maria del Carmen Rodriguez-Hernandez et al. (2020) also propose the content-based Movie recommendation system but compare with several recommendations approaches such as content-based recommender based on vector space models, a deep learning and content-based recommendation approach using Bidirectional Recurrent Neural Networks (BRNNs), and a semantic-aware content-based recommendation model using BERT. The models are evaluated with precision, recall, F1 measure, NDCG@k, (Mean Absolute Error) MAE, and Root Mean Square Error (RMSE). The semantic-aware content-based recommendation model using BERT has shown the best performance to alleviate the cold start problem. The author plans to include more structured features and different external knowledge sources in BERT. Upgrade the size of MovieLens datasets. The author also decides to implement the BRNNs model with different pre-trained embedding models.

Mohadeseh Kaviani et al. (2020) proposed a novel hashtag recommendation using BERT and neural networks. In this work, the hashtags vectors that generate in preprocessing are embedded by using BERT and the vectors are calculated by using cosine similarity to find a similar hashtag to recommend a similar novel to the user. Accuracy, Precision, Recall, and F-1 measure metrics are used to evaluate the proposed model in this paper. Based on the evaluation metrics, BERT has outperformed 3 other embedding methods which are LDA, SVM, and Topic translation model (TTM).

Ram Krishn Mishra et al. (2019) have proposed a hotel recommendation system based on the reviews of the hotels. In this paper, TF-IDF is also used to vectorize the reviews of the hotels and Cosine Similarity is calculated for the similarity between the hotels. Both of these methods are used to provide options more widely on the hotels' recommendations. The author decides to extend the proposed method with other algorithms to improve the accuracy of the systems in the future.

C.P. Patidar et al. (2020) also have proposed a recommendation system that recommends news to the user. Naive Bayes classification is used in this proposed method to classify the data into various categories before being vectorized by TF-IDF. During the evaluation, the precision,

recall, and f-score will be calculated to measure the performance of the proposed system. The result of this system is that TF-IDF could decide the relevant news by finding the similarity factor.

Braja Gopal Patra et al. (2020) propose a content-based literature recommendation system for the medical field used to categorize the literature of different diseases into a correct field. There are multiple embedding methods proposed by the author such as TF-IDF, BM25, Latent Semantic Analysis, Latent Dirichlet Allocation, word2vec, and doc2vec. Cosine similarity is used as the similarity metrics to calculate the similarity between updated paper and paper in the dataset. After several evaluations, BM25 is applied as the top-reforming literature recommendation technique in this paper. The future aim of the author is to incorporate novel retrieval methods, extend the manual evaluation to other datasets and extend this recommendation to other datasets.

G. Sunandana et al. (2021) have built an enhanced content-based filtering movies recommendation system using TF-IDF vectorization. In this recommendation system, the system only deals with the active user to suggest movies. The system will use cosine similarity to evaluate the similarity between the movies' features and use the IMDB formula to get the popularity of the movies. The author decides to build the recommendation system using a hybrid approach to improve the result of the system.

Luong Vuong Nguyen et al. (2020) propose a content-based collaborative approach recommendation system to solve the problem of cold start from a collaborative filtering recommendation system. The author applies the Jaccard coefficient index to convert the extracted features of the movies from the OMDb API database. Besides, Word2Vec is the word embedding model used in this proposed system and the similarity of the movies is calculated by soft cosine measure instead of using cosine similarity. During the evaluation, Accuracy, Precision, Recall, and F1 are used as the evaluation matrices to compare the KNN, Correlation-based Items Similarity (CIS). and Word2Vec. The result of the comparison is that Word2vec has the highest accuracy compared with the other two methods. The author also decides to apply the proposed

similarity measure to other types of recommendation systems and measure the similarity by using other metrics.

Yuyangzi Fu et al. (2020) present a novel approach for item-based collaborative filtering using BERT. The proposed method can solve the common issue of the collaborative filtering recommendation: cold start and “more of the same" recommended content. Then, the proposed method uses the item’s title tokens as content to address the cold start problem. During the evaluation, Bidirectional Long short-term memory (Bi-LSTM), BERT Masked Language Model (BERTbase w/o MLM), and BERT are compared together and the BERT shows the best performance.

Yeo Chan Yoon et al. (2018) have proposed a movie recommendation system using the Word2Vec algorithm. Word2Vec is used to embed metadata information of the movie and the embedding data is used to compare with the historical data from users to recommend some suitable movies based on the users’ preference. Lastly, the result of the proposed method is Word2Vec is the most effective method compared with two baseline methods: Item2vec and SVD method.

Table 2.2: The summary of the word embedding literature review.

Author	**Year**	**Method**	**Dataset**	**Conclusion**
Joanna Cristy Patty et al.	2018	TF-IDF	Own collected dataset	TF-IDF succeed to help the recommendation system to recommend cosmetics.
Tessy Badriyah et al.	2018	TF-IDF	Own collected dataset	The system recommends property by following the property advertisement data from the users successfully using TF-IDF.
Yeo Chan Yoon et al.	2018	Word2Vec, SVD, Item2vec	MovieLens 10M	Word2Vec shows higher performance than the other baseline method.
Ram Krishn Mishra et al.	2019	TF-IDF	Own collected dataset	TF-IDF success to recommend similar types of hotels based on the reviews of different users from the dataset.

Xusong Chen et al.	2019	Bidirectional Encoder Representation from Transformer(BERT),CBF, NARM	Dataset is provided by the challenge organizers	BERT shows the best performance compared with CBF and NARM. The author plans to study the BERT model with mixed content features and embedding features.
Yuyangzi Fu et al.	2020	BERT, LSTM	Dataset is from an e-commerce website data,	BERT model greatly outperforms the LSTM-based model.
Maria del Carmen Rodriguez-Hernandex et al.	2020	TF-IDF, BRNNs, BERT	Movielens, DBook	The results show that the recommender using BERT presents the best performance to alleviate the cold start problem.
Braja Gopal Patra et al.	2020	TF-IDF, BM25, LSA, LDA, word2vec, doc2vec	GEO Dataset	The BM25 applied the top-reforming literature recommendation technique in this paper.
C.P. Patidar et al.	2020	TF-IDF	BBC Dataset	TF-IDF could help to find the similarity factor and recommend the most similar news in the proposed system.
Luong Vuong Nguyen	2020	Word2Vec, KNN, CIS	OMDb API, own collected dataset	Word2Vec always archived higher performance than KNN and CIS.
Mohadeseh Kaviani et al.	2020	BERT, LDA, SVM, TTM	Own collected dataset	BERT is compared with other methods and it outperforms other methods.
Hao Chang et al.	2021	TF-IDF,Word2Vec	Own collected dataset	The recommendation system can solve the recommendation problem when the user has no clear intention and predict the user's behavior based on the comments using TF-IDF and Word2Vec.
G.Sunandana et al.	2021	TF-IDF	Different datasets from Kaggle	The recommendation system uses the information of the active users to recommend successfully using TF-IDF.

2.2.1 Sentiment Analysis

S.M. Asiful Huda et al. (2019) built a model for Bangladesh restaurants using natural language processing and machine learning. The model is built to help the restaurants to investigate the preferences of customers based on their reviews. To build the sentiment analysis model, TF-IDF and Count vectorizer are used. Then, there are 4 classification models used in both sentiment analysis models. As a result, Naive Bayes performs the CountVectorizer model better and SVM performs the TF-IDF model better.

R.M. Gomathi et al. (2019) propose a machine-learning algorithm to solve the issue of the personalized Restaurant selection relying on the tripadvisor.com search data. NLP, a machine learning technique to understand the meaning of human language is used in this work to find the aspect and sentiments of the user comments. During the evaluation, probabilistic neural network (PNN), Back Propagation Neural Network (BPN), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA) is used to compare with NLP and the final result is NLP has the highest accuracy which is 92.45%.

Alia Karim Abdul Hassan et al. (2017) have proposed a review sentiment analysis for a collaborative recommender system to solve the common problem of the collaborative filtering recommendation: data sparsity. To find a better classification and regression method for the sentiment analysis, the author has compared three methods in the Yelp restaurant dataset, IMDB dataset, and Arabic Qaym.com dataset which are Naive Bayes classification, Logistic regression, and Decision tree. Logistic regression and Naive Bayes have performed better in these three datasets but Logistic regression can not handle well in the small dataset like the Yelp restaurant dataset. Besides, sentiment analysis is implemented by using NLTK which is the library of natural language processing. In the future, the author decides to recommend based on the items taken from reviews.

Latent Dirichlet allocation (LDA) is a popular technique for semantic analysis in topic modeling and text mining. Hamed Jelodar et al. (2019) present recommendation systems and

applications based on LDA. The authors use LDA and Gibbs sampling algorithms to evaluate ISWC and WWW conference publications from the DBLP website. This study approved that the recommendation systems based on LDA will be very effective to understand the behaviors of the people to build a recommendation system.

Nimish Kapoor et al. (2020) have proposed a movie recommendation system using NLP tools. SVM is used to build the sentiment analysis system of this recommendation system to predict positive and negative sentiment from user's movie reviews and rate the movies based on the review score. SVM model has reached about 85% of accuracy during the evaluation which is a very high performance. Then, the author recommends this text review analysis-based system can be implemented with other recommendation systems as well.

Vishwa Shrirame et al. (2020) proposed a collaborative filtering recommendation system using sentiment analysis and neural networks. This model is proposed to understand and extract consumer opinions to increase the user experiences of consumers. Various models are used in sentiment analysis: KNN, Random Forest, Naive Bayes, SVM, Gradient Boosting Classifier, and RNN. Then, the Naive Bayes classifier has the best performance for the sentiment analysis.

Table 2.3: The summary of the sentiment analysis literature review.

Author	**Year**	**Method**	**Dataset**	**Conclusion**
Alia Karim Abdul Hassan et al.	2017	Naive Bayes, Logistic regression, Decision tree	Yelp restaurant dataset, IMDB reviews dataset, Arabic qaym.com restaurant reviews dataset	Naive Bayes and Logistic regression can predict sentiment analysis well in these three datasets where the sentiment analysis model is trained by the NLTK library.
Hamed Jelodar et al.	2019	LDA, Collaborative Filtering, Gibbs Sampling	ISWC and WWW conferences publication from DBLP website	LDA is very effective to understand the behaviors of the people to build a recommendation system.

S.M. Asiful Huda et al.	2019	Naive Bayes, SVM, Linear Regression, Decision Tree	Own collected dataset	Naive Bayes model performs well for CountVectorizer and SVM model performs well for TF-IDF.
R.M. Gomathi et al.	2019	NLP, PNN, BPN, SVM, LDA	A dataset from Tripadvisor.com	NLP approach obtained the highest accuracy which is 92.45%.
Nimish Kapoor et al.	2020	SVM, TF-IDF, KNN,	A movie dataset (IMDB)	SVM model presents the best performance compared with the other two baseline methods.
Vishwa Shrirame et al.	2020	KNN, Naive Bayes, Random Forest, SVM, Gradient Boosting Classifier, Recurrent Neural Networks	Amazon product reviews dataset	Naive Bayes classifier has the best performance for the sentiment analysis.

Chapter 3 Proposed Framework

In this section, we discuss the overall algorithmic framework for this project which will be performed in phase 1 and phase 2 as shown in Figure 3.1. Phase 1 focuses on restaurant recommendations based on reviews is by comparing the word embedding method after preprocessing the dataset. There are four word embeddings used such as Term Frequency-Inverse Document Frequency (TF-IDF), Bidirectional Encoder Representations from Transformers (BERT), Word2Vec, and CountVectorizer. During the word embedding process, the reviews of restaurants that had been cleaned during preprocessing will be used in the word embedding method. After word embedding, the data will be counted by cosine similarity to calculate the similarity between the restaurant. Then, the recommended restaurant will be shown based on the cuisines from the input. Evaluation will be carried out using the supervised proxy problem by training classification models. The Precision, Recall, Accuracy, and F1 score of the classification models will be calculated and compared which will be explained in the evaluation section.

Next, the framework of phase 2 which recommends food based on restaurants will be conducted which performs food named entity recognition on reviews from restaurants followed by sentiment analysis on it. Sentiment analysis and food named entity recognition are used to increase the completion of the recommendation model. Then, NLTK and spaCy will be used to find out and work with sentiment analysis on the food entities in the restaurants' reviews.

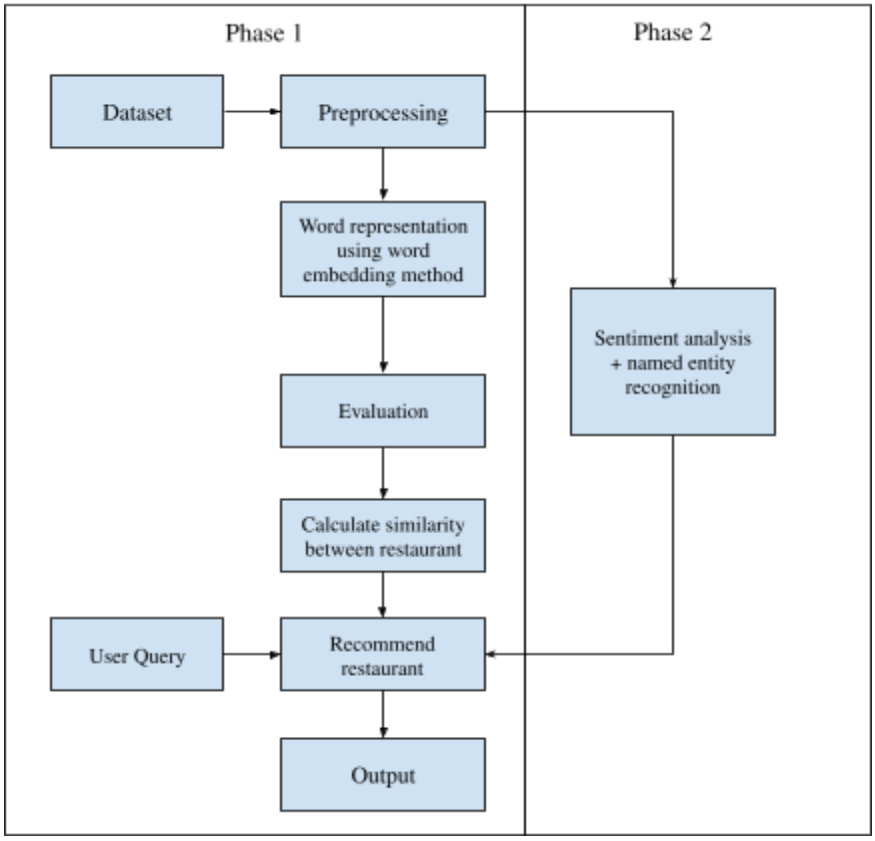

Figure 3.1: Proposed algorithmic Framework

3.1 Word Embedding method

In this section, the word embedding methods that are used in the proposed system will be explained which are TF-IDF, BERT, Word2Vec, and CountVectorizer.

3.1.1 Term Frequency- Inverse Document Frequency (TF-IDF)

TF-IDF is a weighting technique for information retrieval and text mining that shows the importance of a word for a document. TF-IDF consists of Term Frequency (TF) and Inverse Document Frequency (IDF). TF represents the frequency of a word appearing in the document, but IDF is the word with high frequency in the other documents that will become less important.

3.2.1.1 Term Frequency

TF is about counting the number of times a word appears in a document. The more the word appears in the document, the more important the word is in the document. But in this case, a very common problem in calculating TF is the words which are the most common words in the document like “is, am, are, was, were” will be given with a very high TF score based on the idea of the TF. So, these kinds of words which are called stop-words will be removed in the preprocessing step of the documents. After this step, the word that is important in this whole document can be represented through the calculation of TF more easily. Below shows that the formula of the TF :

$$TF(t, d) = \frac{N(t_i, d_j)}{N(d_j)}$$

t_i is the words that appear in the document d_j. $N(t_i, d_j)$ is the number of times that t_i appear in d_j. The total number of words in the document is represented as $N(d_j)$ in this formula.

3.1.1.2 Inverse Document Frequency

IDF helps to measure the informativeness of the word in a document. During the calculation of IDF, the most appearing words will be given with a very low value. It is because the idea of IDF is the more the frequency of the word appears in every document, the less important the word in the particular document.

$$IDF(t_i) = log\frac{N(d)}{N(t_i, d)}$$

$N(d)$ is the number of documents. $N(t_i, d)$ is the number of times that t_i appear in d.

Example calculation of TF-IDF:

Assume document contains100 words and the *bread* appears 10 times in the document.

$$TF(t,d) = \frac{10}{100}$$
$$= 0.1$$

Assume have 10 million documents and the *bread* appears in one thousand times.

$$IDF(t_i) = log\frac{10{,}000{,}000}{1000}$$
$$= 4$$

TF-IDF weight:

$$TF \times IDF = 0.1 \times 4$$
$$= 0.4$$

3.1.2 CountVectorizer

CountVectorizer is a word embedding tool that will transform the words of the document into vectors based on the frequency of each word that appears in the document. A matrix will be created by CountVectorizer by forming each unique word into a column of the matrix, and the document will be the row of the matrix. The word frequency in the document will be counted and recorded into the matrix. After recording the frequency of the words, each row of the matrix will be transformed into a vector.

Below is the example of the CountVectorizer:

Document 1: “A very nice place for board game lovers”

Document 2: “Nice place for cafe lovers”

Table 3.1: Sequence of the word in each corpus.

document/word	**board**	**cafe**	**for**	**game**	**lovers**	**nice**	**place**	**Very**
1	1	0	1	1	1	1	1	1
2	0	1	1	0	1	1	1	0

3.1.3 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a transformer-based machine learning technique for NLP pre-training which is developed by GOOGLE. It was trained on Masked Language Modeling (MLM) and Next sentence prediction (NSP). The MLM is an idea that masks 15% of tokens and BERT was trained to predict them from the context. The idea of NSP is BERT was trained to predict if a chosen next sentence has appeared. BERT learns contextual embeddings for words after pretraining. BERT relies on a transformer, but it only needs the encoder part since its goal is to generate a language representation model. BERT needs the input to be massaged and decorated with some extra metadata before processing. After that, the input to the encoder for BERT is a sequence of tokens, which are first converted into vectors and then processed in the neural network.

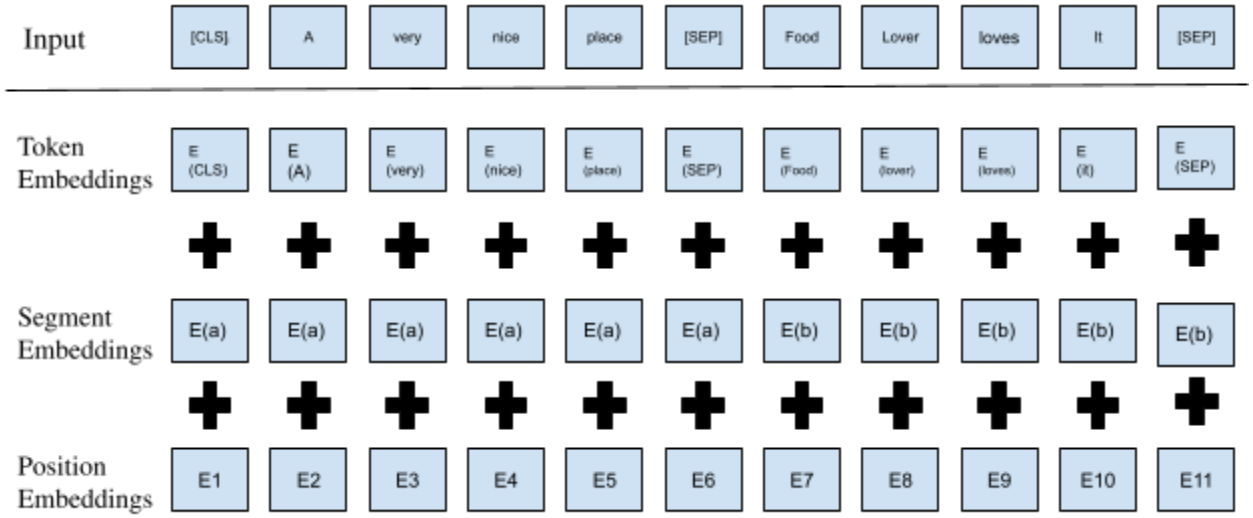

Figure 3.2: The input representation for BERT

Token embeddings: A [CLS] token is added to the input word tokens at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence.

Segment embeddings: A marker indicating Sentence (a) or Sentence (b) is added to every token to allow the encoder to differentiate the sentences.

Positional embeddings: Each token has been added A positional embedding to indicate its position in the sentence.

3.1.4 Word2Vec

Word2Vec is a popular word embedding technique using a shallow neural network. It can be built using Skip Gram and Common Bag Of Words (CBOW).

3.1.4.1 CBOW Model

CBOW model takes the context of each word as the input and predicts the word corresponding to the context.

Example sentence: “A very nice place for board game lovers”

The input word into the neural network is ‘nice’. In the network, the model will try to predict a target word, ‘place’ using a context input word, ‘nice’. One Hot Encoding will be used on the input word to measure the output error compared to the One Hot Encoding of the target word, ‘place’. The vector representation of the target word will learn from this predicting process.

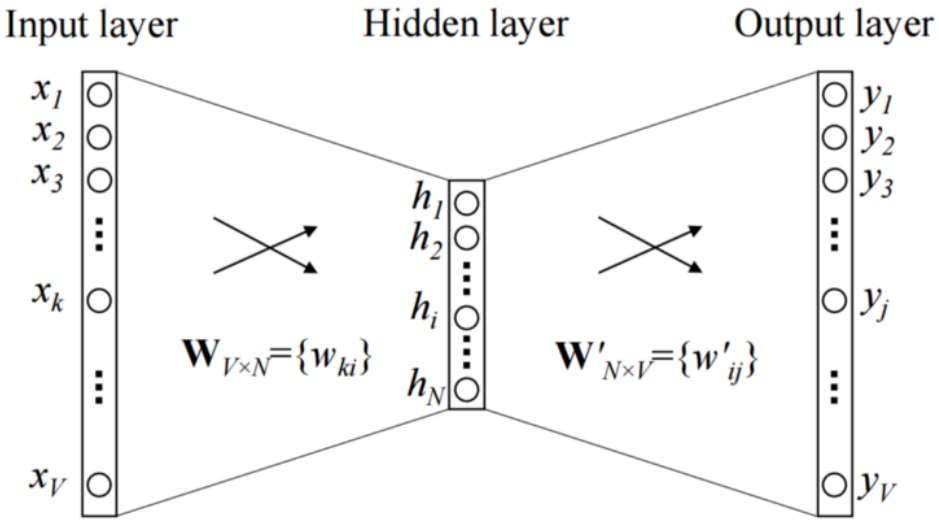

Figure 3.3: The architecture of the CBOW model.

The input word is a one-hot encoded vector of size V and the hidden layer contains N neurons. V length vector with the elements being the softmax values is the output. $w_{V\times N}$ is the weight matrix that maps the input (x) to the hidden layer. Besides, $w'_{N\times V}$ is the weight matrix that maps the hidden layer outputs to the output layer.

3.1.4.2 Skip Gram

Skip Gram is a model just looks like the CBOW model just got flipped. It can use the target word to predict the context. During this process, we produce the representation same with the CBOW model. In the Skip Gram model, we will input the target word into the Neural Network first. Then, the model will give the outputs C probability distribution. We will get C probability distributions of V probabilities which one for each word for each context position. Below is the architecture of the Skip Gram.

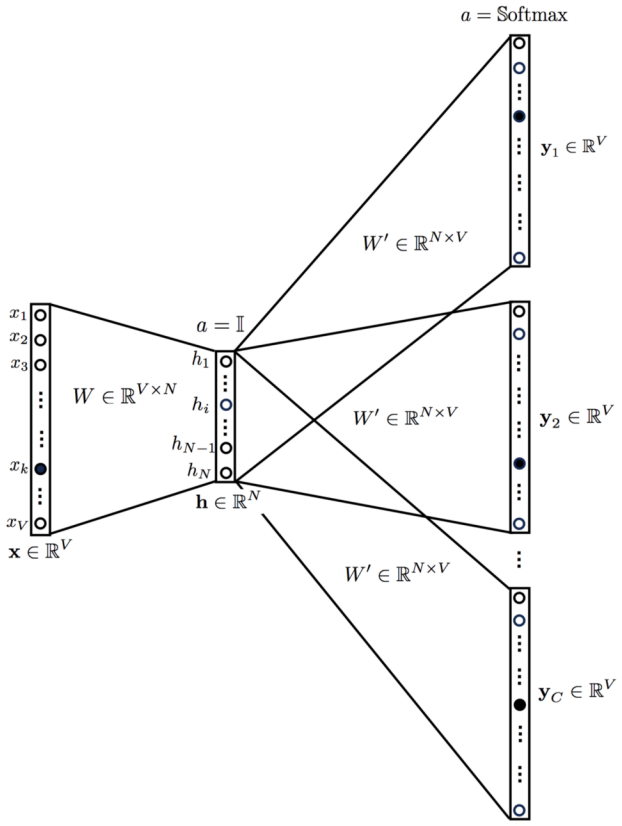

Figure 3.4: The architecture of the Skip Gram model.

3.1.5 Similarity Matrix

The objective of a restaurant recommendation system is to calculate the similarity between the restaurants. In this project, each restaurant's reviews will be represented by a vector. To calculate the similarity of the restaurants, the Cosine Similarity matrix will be used.

3.1.5.1 Cosine similarity

Cosine Similarity is one of the popular measurements to calculate document similarity. It is measured by the cosine of the angle between two vectors to determine if both vectors are pointing in the same direction. The range of the cosine similarity vector is between 0 and 1. 0 means both vectors are in different directions which have no relationship, but 1 means they are in the same direction which is totally the same. Below is the equation of Cosine Similarity.

$$similarity = cos(\theta) = \frac{A - B}{||A|| \, ||B||}$$

$$= \frac{\sum_{n}^{i=1} A_i \times B_i}{\sqrt{\sum_{n}^{i=1} (A_i)^2} \times \sqrt{\sum_{n}^{i=1} (B_i)^2}}$$

Example calculation of the Cosine Similarity:

Example sentences:
Document 1: “A very nice place for board game lovers”
Document 2: “Nice place for cafe lovers”

First step: split the word of both sentences.

Document 1: A, very, nice, place, for, board, game, lovers
Document 2: Nice, place, for, cafe, lovers

Second step: List all the words.

Words: A, very, nice, place, for, board, game, cafe, lovers

Third step: Count the frequency of the words.

Document 1: A: 1, very: 1, nice: 1, place: 1, for: 1, board: 1, game: 1, cafe: 0, lovers: 1
Document 2: A: 0, very: 0, nice: 1, place: 1, for: 1, board: 0, game: 0, cafe: 1, lovers: 1

Fourth step: word frequency vector.

Document 1: [1, 1, 1, 1, 1, 1, 1, 0, 1]
Document 2: [0, 0, 1, 1, 1, 0, 0, 1, 0]

Fifth step: Calculate the cosine of both vectors.

$$cos(\theta) = \frac{1\times 0 + 1\times 0 + 1\times 1 + 1\times 1 + 1\times 1 + 1\times 0 + 1\times 0 + 0\times 1 + 1\times 0}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2+0^2+1^2}\times\sqrt{0^2+0^2+1^2+1^2+1^2+0^2+0^2+1^2+0^2}}$$

The value of similarity of both sentences is 0.6213 which means that the two sentences are slightly similar.

3.2 Evaluation

According to Alexander Geiger (2021), machine learning methods need large amounts of data, but labeled data which is also called benchmark data for a certain use case is rare. Since we do not have labeled data, we will use the unsupervised machine learning methods because it does not require any benchmark data for the training. However, evaluating the performance of unsupervised learning models is more complicated than supervised learning since supervised learning has labeled data that can be used as a target measure but unsupervised learning is not. In this case, one approach is proposed to help evaluate unsupervised learning which is the Supervised proxy problem (Alexander Geiger, 2021)

3.2.1 Supervised proxy problem

In this approach, it trains a related supervised learning algorithm to predict the features of the data which are defined as the target variable using other variables as the input from the data. The performance of this algorithm is used to estimate the performance of word embedding representing the information from the data. So, we will train supervised learning algorithms which are Random Forest, Naive Bayes, and K nearest neighbor (KNN) for the prediction of the proposed evaluation method. To train the classification model, the review vectors from word embedding methods will be split into 70% of training data and 30% of test data. The classification model needs to predict the rating of each testing data after training.

Random Forest Classification: A supervised learning algorithm that consists of many decision trees by trained with the 'bagging' method which is the combination of learning models to increase the overall result. The reason that we use Random Forest Classification is it can handle large datasets efficiently. Besides, the main reason is it provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

Naive Bayes Classification: A classification technique that can assume the existence of a particular feature in a class is not related to the existence of any other feature by using the Bayes theorem to calculate the posterior probability to predict the correct result. Then, its advantage is it can perform well in multi-class prediction and predict the class of the test dataset easily and efficiently.

K Nearest Neighbor: A type of supervised learning algorithm that can be used for regression and classification. It can use to predict the correct class for the test data by calculating the distance between test and train data. K number of points is selected before predicting the data to find the class that has the highest probability. Its main advantage is that it can also be used for multiclass classification.

3.2.2 Precision, Recall, F1 score, and Accuracy

A classification report will come out after classification, and the results will be calculated by following the confusion matrix table (Table 3.2).

Table 3.2: confusion matrix table.

Actual/Predicted	**False (0)**	**True (1)**
False (0)	TN	FP
True (1)	FN	TP

True Negative (TN) means the actual value was False, and the predicted value was False. False Positive (FP) means the actual value was False, and the predicted value was True which is named as a Type I Error. Then, False Negative (FN) means that the actual value was True, and the predicted value was False which is named as a Type II error. Last is the True Positive (TP), which means that the actual value was True, and the predicted value was True.

3.2.2.1 Precision

Precision is the ratio of correctly predicted positive observations to all predicted positive observations.

$$Precision = \frac{tp}{tp + fp}$$

3.2.2.2 Recall

Recall is the ratio of correctly predicted positive observation to all observations in the actual class.

$$Recall = \frac{tp}{tp + fn}$$

3.2.2.3 F1 score

F1 score is the average of Precision and Recall. It takes both false positives and false negatives.

$$F1 = \left(\frac{tp}{recall^{-1} + precision^{-1}}\right)$$

$$= 2 \times \frac{precision \times recall}{precision + recall}$$

3.2.2.4 Accuracy

Accuracy is the most intuitive performance measure for the model. It is a ratio of correctly predicted observations to the total observations.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

3.3 Named Entity Recognition

Named Entity Recognition is one of the subfields in Natural Language Processing (NLP). NER is always used in entity chunking, extraction, or identification to identify and categorize the main information from text. Every entity found is classified into a predetermined category. For example, a NER machine learning model might detect the word “Starbucks” in a text and classify it as a “Company”.

Unnamed: 0.1.1	name&num	reviews_list	Entities	FOOD
0	Le Charcoal Xpress, 0	Great food	[]	[]
1	Le Charcoal Xpress, 0	Had alfaham one shawarma plate and falafel	[('one shawarma plate', 'FOOD'), ('falafel', '...	['one shawarma plate', 'falafel']
2	Le Charcoal Xpress, 0	All crash and tastes really good	[]	[]
3	Le Charcoal Xpress, 0	Prime place in hSR right next to MK retail ma...	[('MK', 'PRODUCT')]	[]
4	Le Charcoal Xpress, 0	Bad	[]	[]

Figure 3.5: The example of the food entity from the reviews.

To define the food entity from the reviews of restaurants, we have found a trained Food Named Entity Recognition model from DeepNote.com. This model used spaCy to train and its accuracy mentioned by the author is 94.14% which is a good result for a model. So, we decide to use this pre-trained model to find out all food entities from the reviews of restaurants. In Figure 3.5, the ‘Entities’ column shows entities other than food entities. For example, “MK” in row 4 of the data is mentioned as a “Product” entity. Then, the ‘FOOD’ column has been removed from all the entities other than the food entity.

3.4 Sentiment Analysis

Sentiment Analysis is also one of the subfields in Natural Language Processing which is used to determine whether the text data is positive, negative, or neutral. Sentiment analysis normally is used to perform on textual data to help application monitor the brand and product sentiment from their customer's feedback to understand customer needs and enhances their services.

Unnamed: 0	name&num	reviews_list	sentiment_neg	sentiment_neu	sentiment_pos	sentiment_compound
0	Le Charcoal Xpress, 0	Great food	0.000	0.196	0.804	0.6249
1	Le Charcoal Xpress, 0	Had alfaham, one shawarma plate and falafel	0.000	1.000	0.000	0.0000
2	Le Charcoal Xpress, 0	All crash and tastes really good	0.273	0.404	0.323	0.1263
3	Le Charcoal Xpress, 0	Prime place in hSR right next to MK retail ma...	0.000	1.000	0.000	0.0000
4	Le Charcoal Xpress, 1	Bad	1.000	0.000	0.000	-0.5423

Figure 3.6: The example of the sentiment scores from the reviews.

After defining the food entity from the reviews, we have also calculated the sentiment score from every sentence. Based on Figure 3.6, there are 4 types of sentiment scores: negative, neutral, positive, and compound score. To determine the reviews' sentiment easily, we only use the compound score for the project. So, we decide to define the sentiment of the review by using the standard in Table 3.3.

Table 3.3: Sentiment Analysis table.

Sentiment Score	**Sentiment**
Sentiment Score > 0	Positive
Sentiment Score == 0	Neutral
Sentiment Score < 0	Negative

To define the food entity's sentiment, we decide to find out the pattern of the sentences by using Part of Speech (POS) and combine with the sentiment result from the Table 3.3. Then, we try to find out the sentences' POS with the food entity which is the positive and negative sentiment from the sampling of the dataset and we have only selected the top 5 POS from both sentiments which are shown in Table 3.4. Besides, we have chosen the POS that only appears together with 'JJ' (Adjective) and 'NN' (Nouns) because 'JJ' can represent the sentiment of the text and 'NN' can be represented the food entity from the reviews. After that, these POS selected will use to define the food entity's sentiment from every review.

Table 3.4: Top 5 Parts of Speech in Positive and Negative.

Positive + Food NER	**Negative + Food NER**
'DT', 'NN', 'NN', 'JJ'	'NN', 'PRP', 'VBD', 'JJ', 'NN', 'NN', 'NNS', 'NNP', 'NNS', 'VBD', 'JJ', 'NNS', 'VBD', 'RB', 'JJ', 'NNS', 'VBP', 'NNP', 'NNP'
'DT', 'NN', 'NN', 'RB', 'JJ'	'NNP', 'JJ', 'NN', 'VBN', 'NN', 'NN', 'NN', 'JJ', 'NN', 'VBG'
'DT', 'NN', 'JJ', 'NN'	'NNP', 'NN', 'NNP', 'VBZ', 'JJ', 'VB', 'NNS', 'VB'
'DT', 'NN', 'NN', 'NN', 'JJ'	'PRP', 'VBP', 'JJ', 'NNS', 'VBP'
'JJ', 'NN', 'NNS'	'IN', 'NNS', 'JJ', 'NN', 'NNS', 'VBP', 'JJ', 'NN', 'NNS'

Chapter 4 Results and Discussions

In this section, we will present our investigation before starting to build the framework. For example, a survey analysis and exploratory data analysis. Besides, we will also discuss the recommendation and evaluation result of the proposed method.

4.2 Survey Analysis

To investigate the ideas of people nowadays on the recommendation system, we have created a very simple survey and sent it to 50 people to fill it up. This survey is made up of 11 questions which are asking about personal information like age and gender etc, ideas on recommendation system, and impressions on using the existing online system which implements the recommendation system.

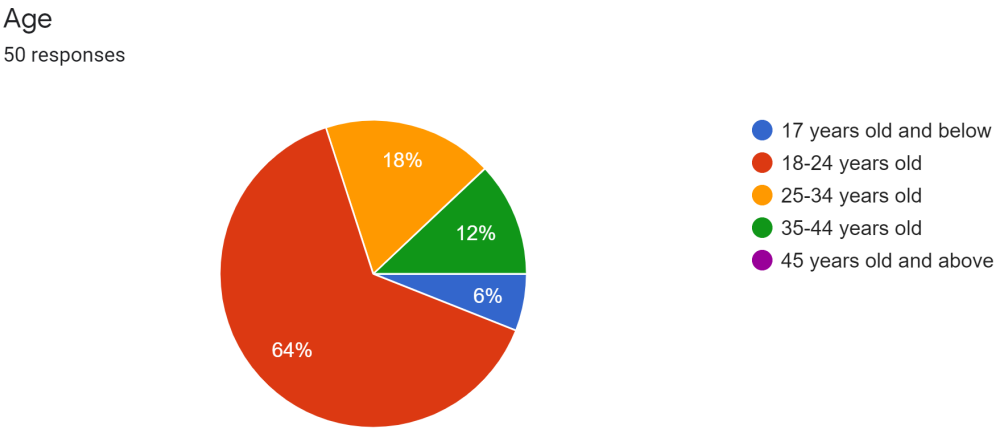

Figure 4.1: Age range of respondents.

The pie chart above (Figure 4.1) shows the age of the respondents. Based on the chart, the age range of the respondents is between 17 to 44 years old. The 18 to 24 years old has the highest number of respondents which is about 64% of respondents and followed by 25 to 34 years old, 35 to 44 years old, and 17 years old and below with 18%, 12%, and 6% of respondents.

Besides, the next question is asking about the gender of the respondents. In the pie chart below (Figure 4.2), there are about 52% of male respondents and 48% of respondents which are nearly equal to each other.

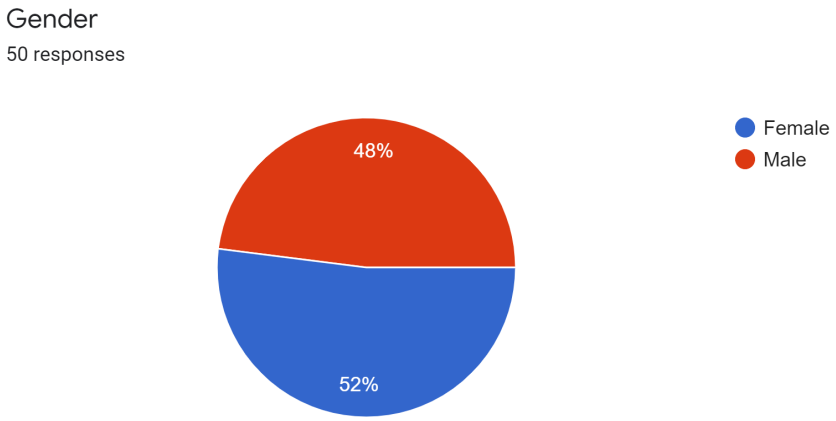

Figure 4.2: Gender of the respondents.

The pie chart in Figure 4.3 shows the result of the employment of the respondents. Based on the age range of the respondents, most of the respondents are 18 to 24 years old which this age range of respondents mostly are students. So, this pie chart shows that 64% of the respondents are students and followed by 22% of the respondents are employed for wages. Then, other respondents are homemakers, out of work, and self-employed.

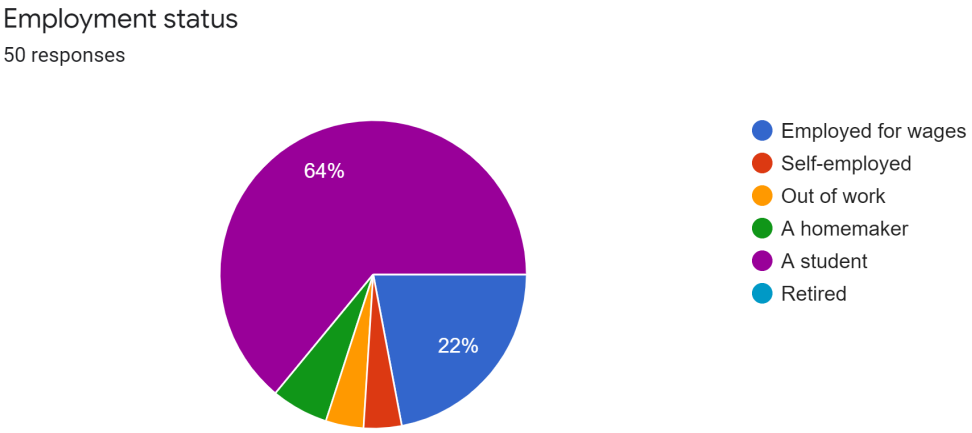

Figure 4.3: Employment status of respondents.

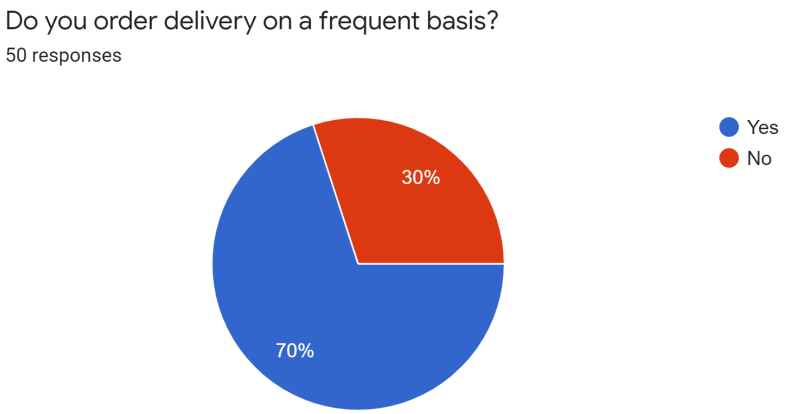

Figure 4.4: The respondents order food delivery frequently or not frequently.

The pie chart in Figure 4.4 shows the results of the respondents about whether the respondents order food delivery frequently or not. The chart shows that 70% of the respondents which is a very high percentage of respondents are order delivery frequently, but only 30% of the respondents do not order delivery frequently.

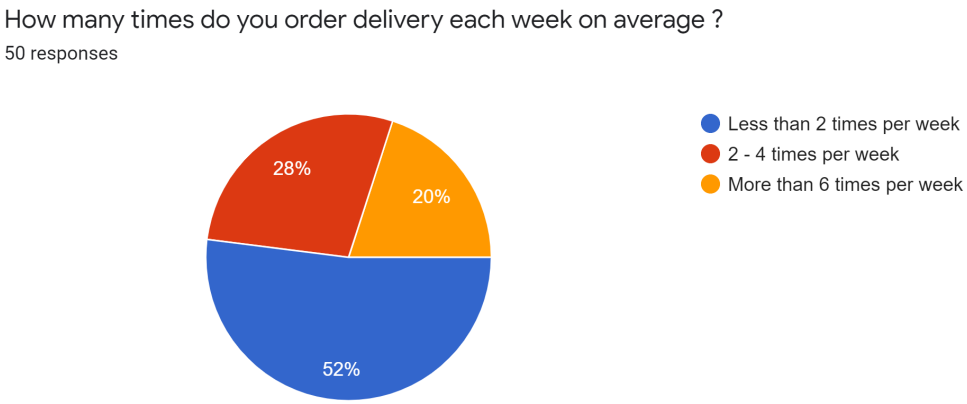

Figure 4.5: The frequency of the respondents ordering food delivery.

Based on the pie chart in Figure 4.5, most of the respondents which are 52% of respondents order food delivery less than twice per week. Less than half of the respondents order food delivery more than twice per week which are 28% of respondents order 2 to 4 times per week and only 20% of respondents order more than 6 times per week.

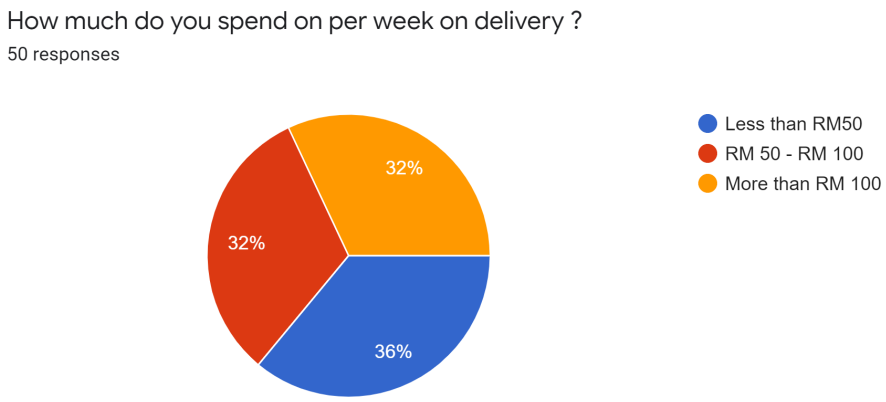

Figure 4.6: The money spent by respondents on food delivery.

The pie chart (Figure 4.6) shows the money spent by respondents on food delivery. In this chart, there are 3 categories of respondents who are spending less than RM50, between RM50 and RM100, and more than RM100. These 3 categories are mostly equal, which are 32% for spending less than RM50, 32% for spending between RM50 and RM100, and 36% for spending more than RM100.

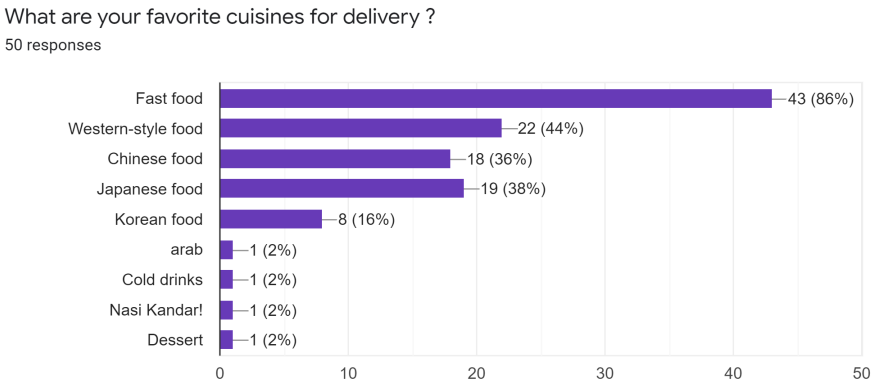

Figure 4.7: The favorite cuisines of respondents on food delivery.

The bar chart in Figure 4.7 shows the favorite cuisines of respondents on food delivery. Based on the bar chart, Fast food is the most popular cuisine for the respondents which have about 86% of respondents love to order it during food delivery. There are some cuisines that are less popular which are Arab food, Cold beverage, Mamak food, and dessert. Only 2% of respondents will order on food delivery. Western-style food, Chinese food, and Japanese food are near to 40% and above of respondents will like, but only 16% of respondents will order Korean food.

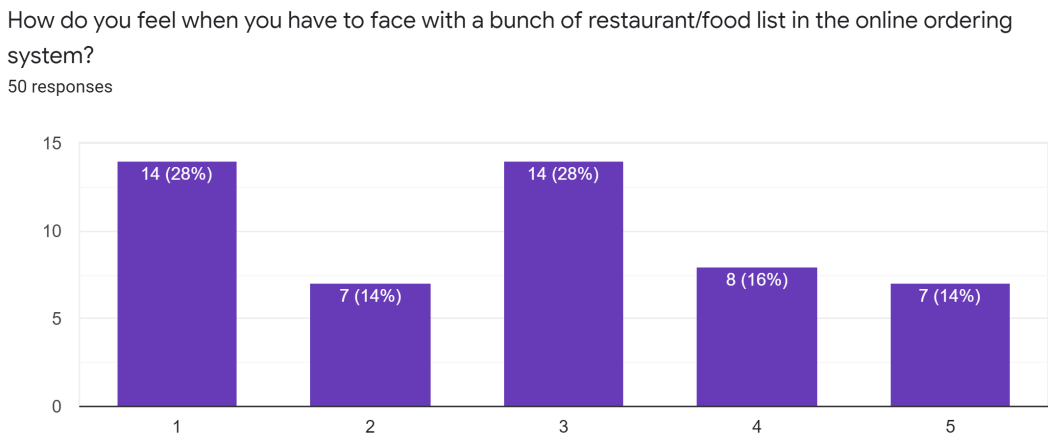

Figure 4.8: the reaction of the respondents facing a brunch of restaurant/food while ordering food.

The bar graph in Figure 4.8 shows the reaction of the respondents while facing a brunch of restaurant/food during food ordering. There are 5 choices for the respondents which are 1 = do not know how to choose, 2 = most of the time will do not know how to choose but less of the time will simply choose one, 3 = sometimes will simply choose one or do not know how to choose, 4 = most of the time will simply choose one but less of the time will do not know how to choose, and 5 = every time will simply choose one. The graph shows that results of 1 and 3 are equal, 28% of respondents choose both of the choices. Besides, other choices: 2, and 5 also have an equal result which is 14% of respondents, but 16% of respondents choose the 4th choice which is most of the time will simply choose one but less of the time will not know how to choose.

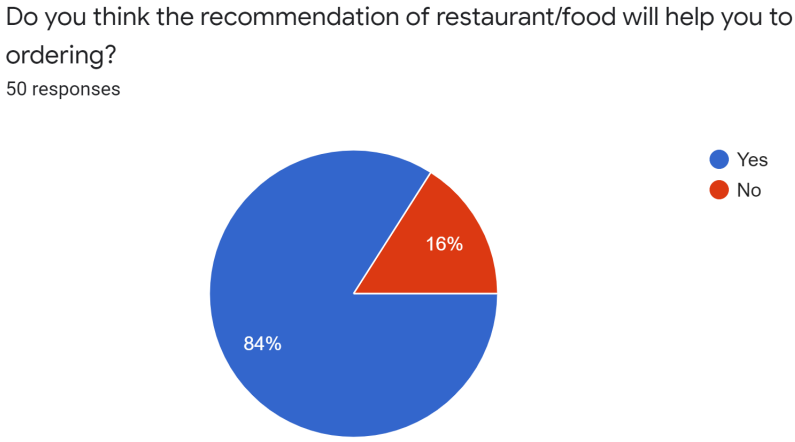

Figure 4.9: The recommendation while ordering brings convenience or not to the respondents.

The question asked in Figure 4.9 shows that most of the respondents which are 84% of respondents think that the recommendation of restaurant/food will help them to be more convenient during ordering in the system, but only 16% of the respondents do not think that the recommendation will bring convenience for them.

Then, the next question (Figure 4.10) is asking about the reason that the recommendation brings convenience to the respondents. Based on the bar chart below, 88% of respondents think that the recommendation system will save their time to find the restaurant or food that the respondents prefer. 60% of respondents think that the recommendation system will help to explore a new restaurant or food easily, but 34% of respondents feel that the recommendation system can help them to prevent wasting money on restaurants or food that they do not prefer.

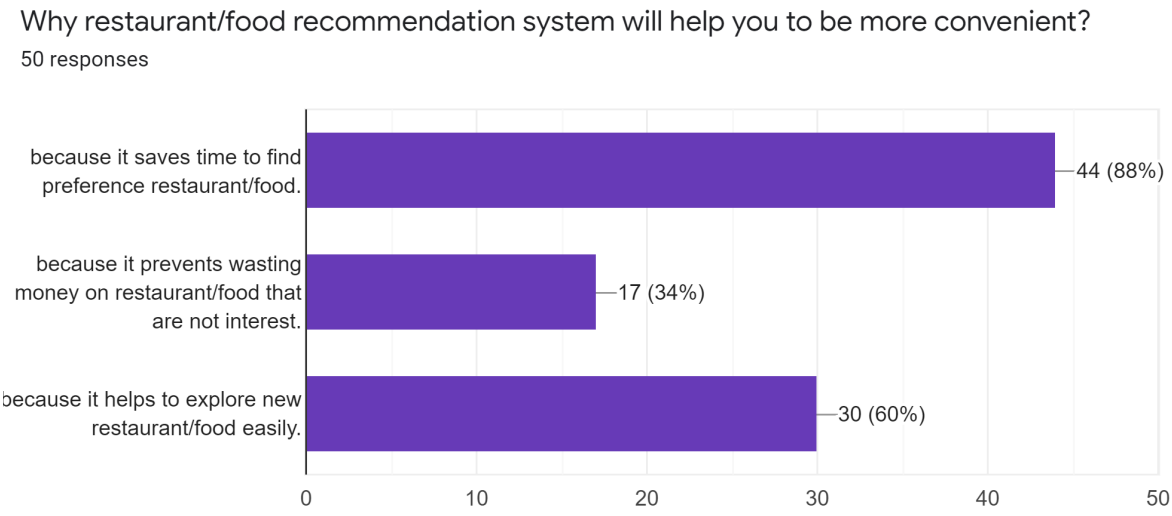

Figure 4.10: The reason for the recommendation system brings convenience to the respondents.

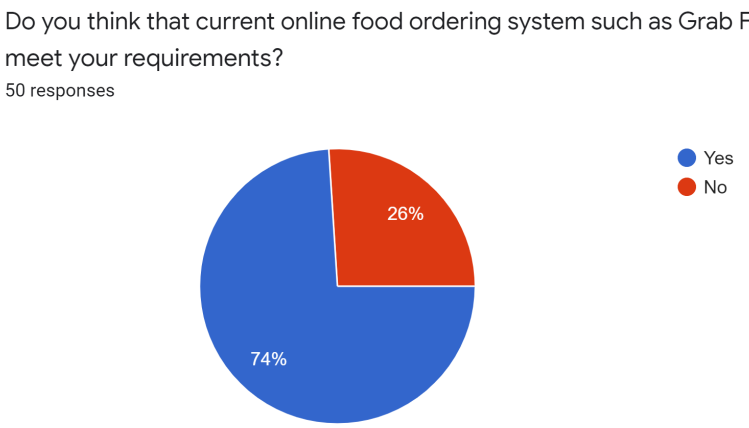

Figure 4.11: The current online food ordering system meets respondents' requirements or not.

Figure 4.11 shows the last question of the survey which is do you think the current online food ordering system such as Grab Food, Food panda, etc. always meets your requirements? The results from the respondents are mostly yes, with 74% of respondents feeling that the systems nowadays always meet their requirements, but only 26% of respondents' requirements do not meet the online food ordering system.

As can be seen from the survey, most people nowadays love to use online ordering systems such as GrabFood, and Foodpanda, etc to order their favorite food without concern about the price of the food. Besides, they also think that the recommendation from the systems is very important which can help them to save money, time or explore new restaurants and food. But, some people are still not satisfied while using the current online ordering system because they feel that their requirements are not met by the systems. In conclusion, the recommendation system is very popular and important for the current online ordering systems, but these systems still need some improvements to meet the users' requirements.

4.3 Dataset

Restaurant dataset from Kaggle which is the Bangalore Restaurant dataset. In this dataset, there are 51717 rows and 17 columns of data which provide us a lot of information about these restaurants very clearly. For example, name, online_order, book_table, rate, rest_type, dish_liked, cuisines, and reviews_list, etc. Especially the reviews_list column can provide complete data to train the proposed model based on the review from the customers in these restaurants.

	url	address	name	online_order	book_table	rate	votes	phone	location	rest_type	dish_liked	cuisines	approx_cost(for two people)	reviews_list	menu_item	listed_in(type)	listed_in(city)
0	https://www.zomato.com/bangalore/jalsa-banasha...	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1/5	775	080 42297555\r\n+91 9743772233	Banashankari	Casual Dining	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese	800	[('Rated 4.0', 'RATED\n A beautiful place to ...	[]	Buffet	Banashankari
1	https://www.zomato.com/bangalore/spice-elephan...	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1/5	787	080 41714161	Banashankari	Casual Dining	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai	800	[('Rated 4.0', 'RATED\n Had been here for din...	[]	Buffet	Banashankari
2	https://www.zomato.com/SanchurroBangalore?cont...	1112, Next to KIMS Medical College, 17th Cross...	San Churro Cafe	Yes	No	3.8/5	918	+91 9663487993	Banashankari	Cafe, Casual Dining	Churros, Cannelloni, Minestrone Soup, Hot Choc...	Cafe, Mexican, Italian	800	[('Rated 3.0', 'RATED\n Ambience is not that ...	[]	Buffet	Banashankari
3	https://www.zomato.com/bangalore/addhuri-udupi...	1st Floor, Annakuteera, 3rd Stage, Banashankar...	Addhuri Udupi Bhojana	No	No	3.7/5	88	+91 9620009302	Banashankari	Quick Bites	Masala Dosa	South Indian, North Indian	300	[('Rated 4.0', 'RATED\n Great food and proper...	[]	Buffet	Banashankari
4	https://www.zomato.com/bangalore/grand-village...	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Grand Village	No	No	3.8/5	166	+91 8026612447\r\n+91 9901210005	Basavanagudi	Casual Dining	Panipuri, Gol Gappe	North Indian, Rajasthani	600	[('Rated 4.0', 'RATED\n Very good restaurant ...	[]	Buffet	Banashankari

Figure 4.12: The first 5 rows of data from the Bangalore Restaurant Dataset.

4.3.1 Pre-processing

Some unused columns will be deleted such as URL, book table, online order, and phone, etc. After deleting the unused columns, duplicates and NaN rows also will be removed to make sure the data is neat and clean. Then, numeric data will also be cleaned such as the rating of restaurants and the rating of the restaurants with multiple outlets will be averaged. The reviews of the restaurant will be cleaned by removing the stopwords, numeric data, URL, and punctuation based on the requirement of the word embedding methods. NLTK library will be used to help on removing the stopwords in the review lists of the restaurants to make sure that the machine can recognize the words correctly during training the Word2Vec, TF-IDF, and CountVectorizer model, but the stopwords in the review list no need to remove while training the BERT model and the process of finding the NER and Sentiment Analysis.

Example of removing stopwords in the review_list column:

Review before removing stopwords

A beautiful place to dine in The interiors take you back to the Mughal era The lightings are just perfect We went there on the occasion of Christmas and so they had only limited items available. But the taste and service was not compromised at all The only complaint is that the breads could have been better Would surely like to come here again

Review after removing stopwords

beautiful place dine interiors take back mughal era lightings perfect went occasion christmas limited items available taste service compromised complaint breads could better would surely like

	address	name	online_order	book_table	rate	votes	location
0	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1	775	Banashankari
1	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1	787	Banashankari
2	1112, Next to KIMS Medical College, 17th Cross...	San Churro Cafe	Yes	No	3.8	918	Banashankari
3	1st Floor, Annakuteera, 3rd Stage, Banashankar...	Addhuri Udupi Bhojana	No	No	3.7	88	Banashankari
4	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Grand Village	No	No	3.8	166	Basavanagudi

rest_type	cuisines	cost	reviews_list	type	city	Mean Rating
Casual Dining	North Indian, Mughlai, Chinese	800.0	A beautiful place to dine in.The interiors ta...	Buffet	Banashankari	3.99
Casual Dining	Chinese, North Indian, Thai	800.0	Had been here for dinner with family. Turned ...	Buffet	Banashankari	3.97
Cafe, Casual Dining	Cafe, Mexican, Italian	800.0	Ambience is not that good enough and it's not...	Buffet	Banashankari	3.58
Quick Bites	South Indian, North Indian	300.0	Great food and proper Karnataka style full me...	Buffet	Banashankari	3.45
Casual Dining	North Indian, Rajasthani	600.0	Very good restaurant in neighbourhood. Buffet...	Buffet	Banashankari	3.58

Figure 4.13: The dataset after data cleaning is done.

4.3.2 Exploratory Data Analysis

In this section, an exploratory data analysis of the dataset proceeded to find out the relationship between the data.

4.3.2.1 location of restaurants

Based on the map below, all of the locations of the restaurants are mostly near Bangalore, India.

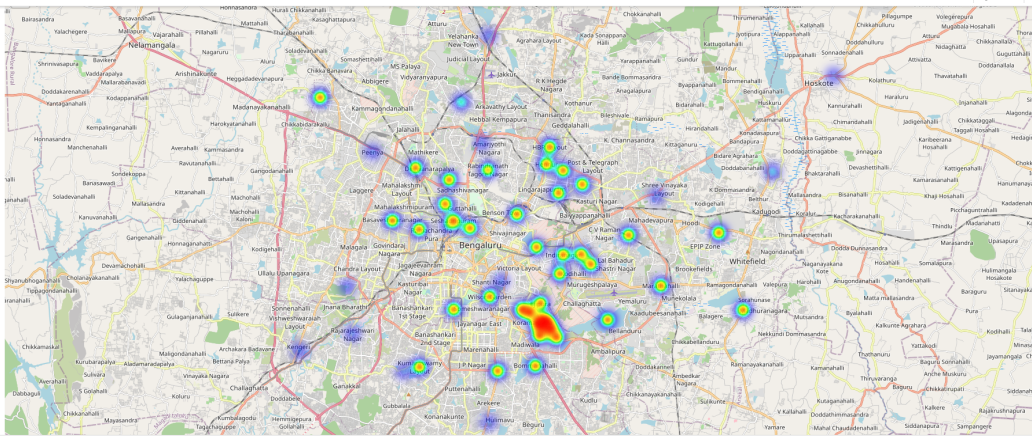

Figure 4.14: Map showing the location of the restaurants.

4.3.2.2 Restaurant types

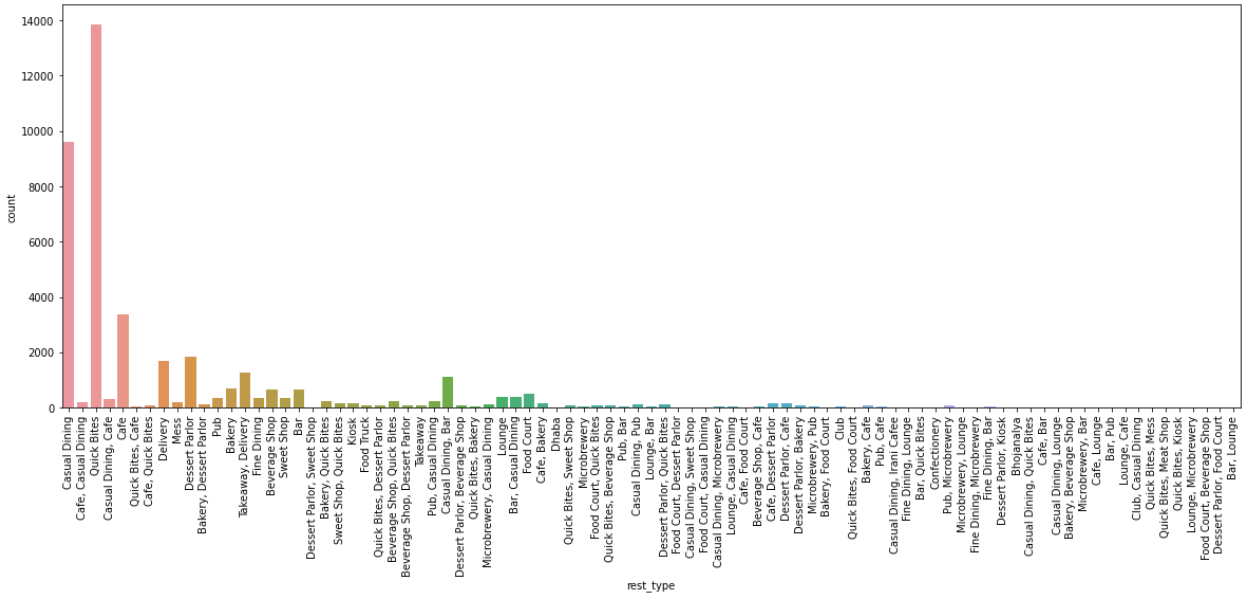

Figure 4.15: Bar chart showing the popular restaurants type.

Based on the bar chart in Figure 4.15, Quick Bites and Casual Dining are the most popular and the second most popular restaurant types which are nearly about 14000 and 10000 restaurants are these restaurants type.

4.3.2.3 Mean rating of every dining type

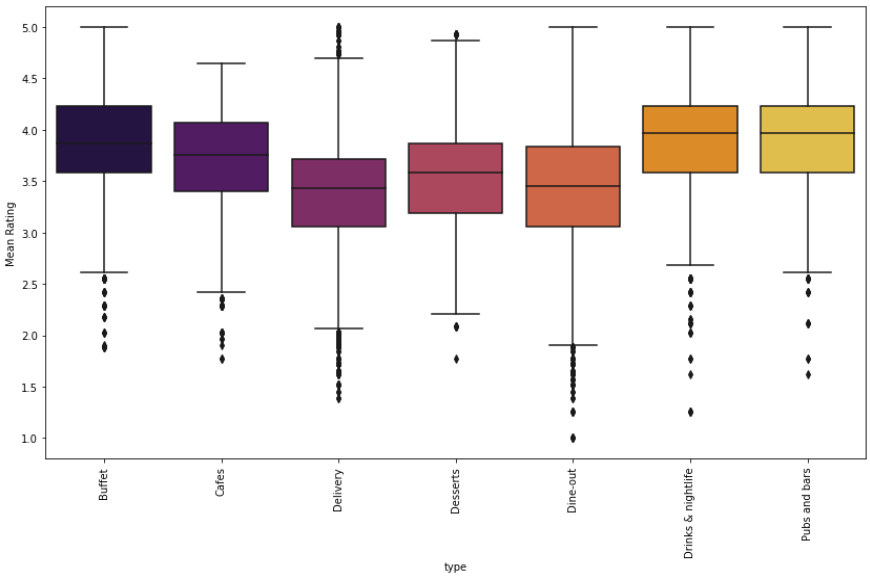

Figure 4.16: Boxplot showing the mean rating of every dining type.

Figure 4.16 is the mean rating of every dining type, it shows that the average of the mean rating range is about 2.0 to 5.0, only the Cafes, Delivery and Dessert does not have the highest mean rating: 5.0 which are only 4.7, 4.8 and 4.9, but it is very near to 5.0. The box of all categories are overlapping and contain a lot of lower outliers especially the Dine out and Delivery category. Only the Desserts and Delivery category have upper outliers.

4.3.2.4 Number of restaurants in every city

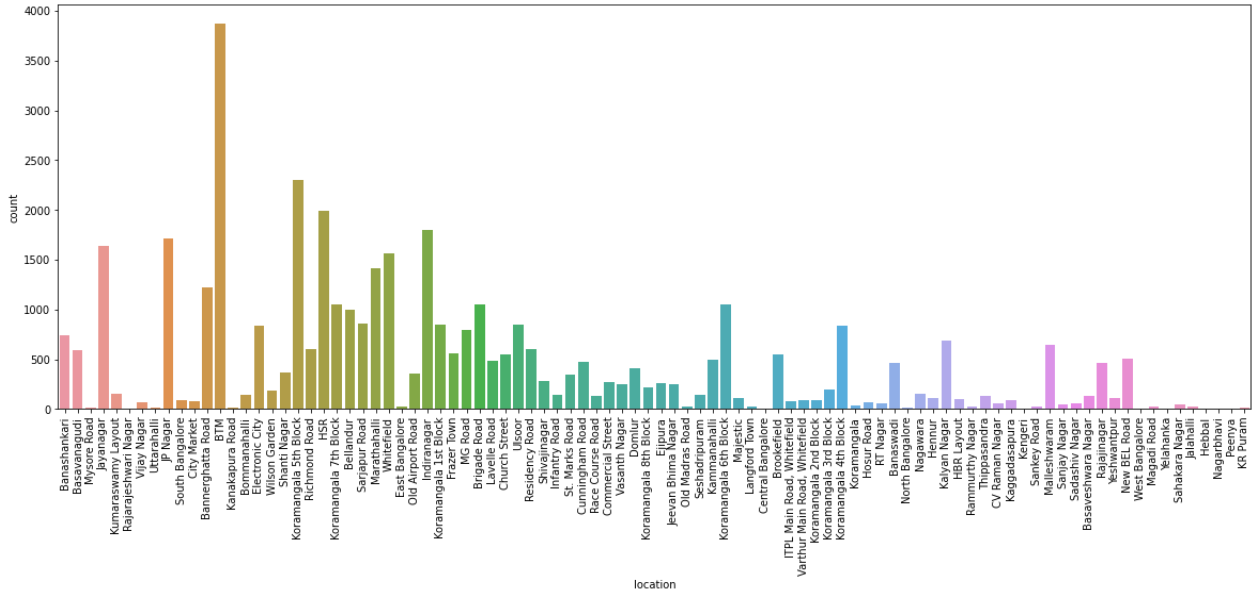

Figure 4.17: Bar chart showing the city of the restaurants.

The bar chart shows the city of the restaurants. Based on the observations, BTM has the highest number of restaurants which is about 3800 restaurants and followed by Koramangala 5th Block which is only 2400 restaurants. Other cities only have below 2000 restaurants.

4.2.2.5 Restaurant's dining types

The bar chart in Figure 4.18 shows the number of restaurants in every dining type. Based on the bar chart below, most of the dining types of the restaurants are below 2000 restaurants, only Delivery and Dine out are more than other dining types which are 20000 and 14000 restaurants.

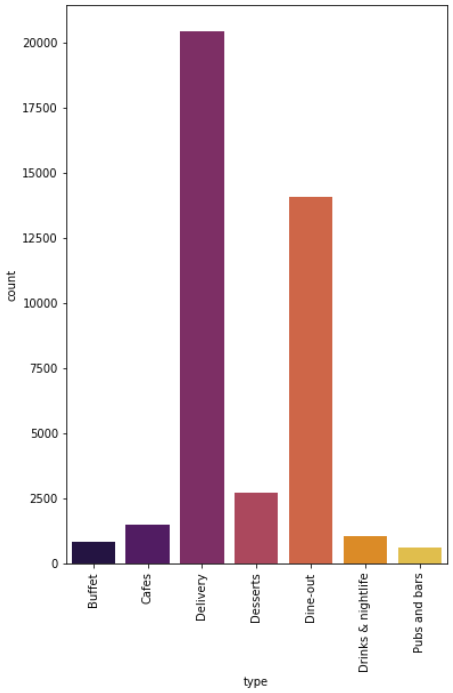

Figure 4.18: Bar chart showing the number of restaurants in every dining type.

4.3.2.6 Cuisine of the restaurants

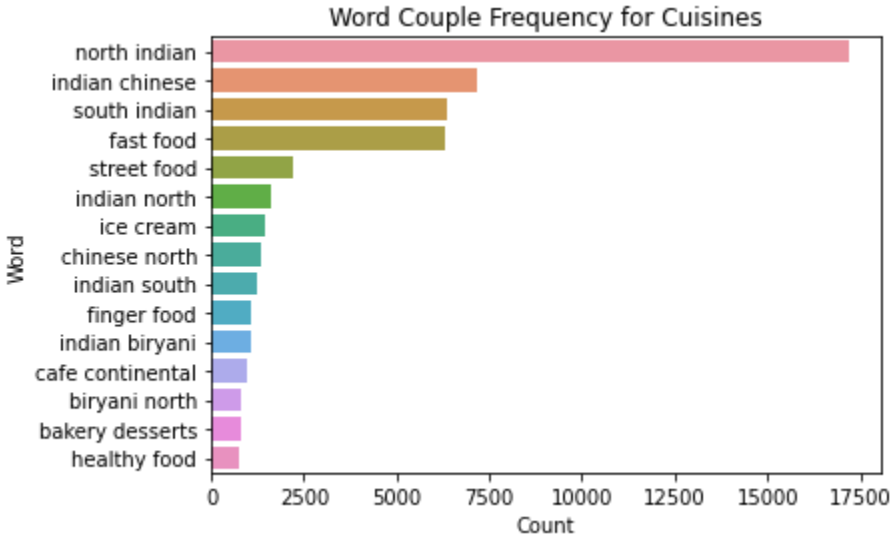

Figure 4.19: Bar chart showing the cuisine of the restaurants.

Based on the bar chart shows the cuisine that has the highest frequency in the restaurants' cuisine. North Indian has the highest frequency which is about 17500 restaurants and followed by Indian Chinese with 7500 restaurants. South Indian and Fast Food has about 7000 restaurants. Besides, other cuisines have only below 2500 restaurants. So, North Indian is the most popular cuisine based on Figure 4.19.

4.4 Recommendation

In this section, we will present and discuss the results of the recommendation from each proposed methods. Figure 4.20 shows the TF-IDF model recommending a list of restaurants that is similar to the North Indian, Chinese, and Biryani restaurants. We can see that the similarity score of the recommended restaurants is only on average 0.55. Although the similarity scores are slightly low, most of the restaurants that are recommended almost meet the requirements where the cuisines of the recommended restaurants are in North Indian, Chinese, and Biryani.

	name	reviews_list	score
North Indian, Chinese	Rangla Punjab	somehow place go place crave plain dal roti sa...	0.578873
North Indian, Biryani, Chinese	Dwaraka Krishna	ordered shahi paneer kalmi kabab tonight place...	0.570915
North Indian, Chinese	Night Food Joint	ordered tandoori chicken naan zomato delivery ...	0.567853
North Indian, Biryani, Chinese	Sri Sai Mango Tree Restaurant	good ordered spl cb friend one veg biryani veg...	0.567039
Chinese	Chung's Chinese Corner	itã xã xã xã xã xã xã xã xs decent place looki...	0.548922
North Indian, Chinese, Biryani	Hyderabad Biryani	awesome biryani day one mostly getting confuse...	0.547527
Chinese	China Bowl	hey cb ordered good old chicken noodle soup la...	0.542146
North Indian, Chinese	Paprica	everything overpriced restaurant looks bad wai...	0.539265
North Indian, Chinese	Owl's Kitchen	pathetic service place order icecream receive ...	0.539082
North Indian, Chinese	3 Spice	ordering food almost years ghee rice grilled c...	0.536691

Figure 4.20: Dataframe shows that the top 10 recommended North Indian, Chinese, Biryani restaurants from TF-IDF model.

Next, we also observe the result from the BERT model (Figure 4.21). The similarity scores between the restaurants are on average 0.9 which are very high but the cuisines of the recommended restaurants are slightly different from the requirement especially the first recommended restaurants: Delight Food's cuisines are Andhra, Chinese, Biryani. Andhra is not one of the required cuisines that we search. Based on the observation, we will conclude that the BERT model can recommend some restaurants that are unexpected.

cuisines	name	reviews_list	score
Andhra, Chinese, Biryani	Delight Food	Came as hunger saviour at night 2 AM..... O...	0.931813
Biryani, North Indian	Biryani Miya	When I felt hungry in the midnight. I always ...	0.929926
North Indian, Chinese, Continental, Biryani, Thai	BHR Foods	Ordered chicken schezwan noodles and chilli c...	0.929849
Fast Food, Rolls, Momos	Pathaan Sir	Visited here last Friday.Tried to have hot mo...	0.929728
North Indian, Chinese	Mak n Vak - Fyn	Ordered chilli chicken. It Was so good and ta...	0.928851
North Indian, Chinese	Punjabi Tadka	Been here multiple times with my friends. If ...	0.928682
Fast Food, Desserts	Sarah's	The chicken burger & veg burger combos were v...	0.927474
North Indian, South Indian, Chinese, Continental	DinePost9	A recent favourite for late night eats... we ...	0.927416
Kerala, Biryani, South Indian, North Indian, C...	Thalassery Restaurant	I had ordered butter naan, tandoori chicken, ...	0.926721
Biryani, Fast Food	Biriyani Adda	Last night I was craving for some good foods ...	0.924374

Figure 4.21: Dataframe shows that the top 10 recommended North Indian, Chinese, Biryani restaurants from the BERT model.

Besides, the next model will be the CountVectorizer model (Figure 4.22). In this model, the average similarity scores are also 0.90 which is also the same as the BERT model but the top model has a 1.0 similarity score which is 'RR Catering', one of the restaurants with the North Indian, Chinese, and Biryani cuisines. The cuisines of the recommended restaurants in the list are also the same as the TF-IDF model is in the searching cuisines: North Indian, Chinese, and Biryani.

	name	reviews_list	score
North Indian, Chinese, Biryani	RR Catering	really felt like good get india always looking...	1.000000
North Indian, Chinese, Biryani	Hyderabad Biryaani Palace	good order delivered loved satisfied ã xã xã x...	0.994480
Biryani, North Indian, Chinese	Hyderabadi Biryani Hub	usually order chicken biryani tasty little bit...	0.847098
North Indian, Chinese	Lovely Knights	food nice mixed north indian flavor thali one ...	0.836869
Chinese	Chungs Pavilion	visiting place since never ever disappointed e...	0.832245
Chinese	Hongkong Noodles	ordered chicken fried rice crazy quantity plat...	0.832158
North Indian, Chinese	Night Food Joint	ordered tandoori chicken naan zomato delivery ...	0.829420
North Indian, Biryani, Chinese	Sri Sai Mango Tree Restaurant	food eatable taste wise items ok lot starters ...	0.826764
North Indian, Chinese	Sufra Restaurant	worst experience sufra restaurant friday night...	0.826256
North Indian, Chinese	Parivar	going yrs still preferred place quality mainta...	0.826253

Figure 4.22: Dataframe shows that the top 10 recommended North Indian, Chinese, Biryani restaurants from the CountVectorizer model.

	name	reviews_list	score
North Indian, Biryani, Chinese	Sri Sai Mango Tree Restaurant	food eatable taste wise items ok lot starters ...	0.991841
North Indian, Chinese	Ghar Ka Chulha	follow instruction specified food packaging ba...	0.991762
North Indian, Biryani, Chinese	Hashtag Kitchen	newly opened place vicinity decided give try p...	0.991650
Biryani, North Indian, Chinese	Jaffa's Biryani	ordered paneer biryani place swiggy packaging ...	0.991617
North Indian, Chinese	Mint and Mustard	good north indian restaurant vv puram excellen...	0.991554
North Indian, Chinese	Hunger Hitman	amazing food deliciously awesome upto expectat...	0.991479
North Indian, Chinese	New Punjabi Dhaba	ordered chicken biriyani taste ok chicken piec...	0.991224
North Indian, Chinese	Hatti Punjab Di	ordered fish curry meal via zomato complaints ...	0.991205
North Indian, Chinese	Wazir's	starting delivery delivery executive polite or...	0.991192
North Indian, Biryani, Chinese	Dwaraka Krishna	ordered shahi paneer kalmi kabab tonight place...	0.991098

Figure 4.23: Dataframe shows that the top 10 recommended North Indian, Chinese, Biryani restaurants from the Word2Vec model.

Last, the list of recommended restaurants from the Word2Vec model is shown in Figure 4.23. We have observed that the similarity scores between the restaurants are very high which are on average 0.99 and the cuisines of the restaurants are in the cuisines that were searched by us.

Figure 4.24: Word Cloud of the reviews' recommended restaurants from each model.

To look through the relationship between the target restaurant, and recommended restaurants, Figure 4.24 has shown the Word Cloud of the reviews of these restaurants in each word embedding method. Word Cloud could display the most prominent or frequent words in the corpus. Therefore, we can observe the relationship between these restaurants.

Based on the Word Cloud of the target restaurant, we found that ‘food’, ‘place’, and ‘chicken’ are the words that most frequently appear in the reviews of the restaurant. If we look to the Word Cloud of the TF-IDF, we can see that ‘food’, ’place’, and ‘chicken’ are also the frequent words in the recommended restaurants, but some words like ‘taste’, and ‘biryani’, etc, the frequent words which are not frequent in reviews of target restaurant. Next, the BERT’s Word Cloud is more similar to the Word Cloud of the target restaurant. The Word Cloud of BERT is also frequent with ‘food’, and ‘chicken’ only. At the same time, the Word Cloud of Word2Vec is frequent with ‘food’ which also appears in the Word Cloud of the target restaurant but also frequent with ‘biryani’ and ‘restaurant’. Lastly, the Word Cloud of the CountVectorizer looks different from the Word Cloud of the target restaurant which is frequent with ‘xã’, the french word that does not appear in other word embedding methods. To summarize, BERT's Word Cloud is the most similar to the target restaurant’s word cloud, but CountVectorizer’s Word Cloud looks not very similar.

4.5 Result

To evaluate the proposed methods used on the recommendation systems, there are 3 classification algorithms built which are Random Forest Classification, Naive Bayes, and K nearest neighbor to predict the rating of restaurants with the vectors from the word embedding algorithms. Then, we will discuss the evaluation results of the classification models on every word embedding method to find out the best performing method.

Table 4.1: Random Forest Evaluation result table of word embedding methods.

Word Embedding Model	**Precision**	**Recall**	**F1**	**Accuracy**
TF-IDF	0.86	0.82	0.81	0.82
Word2Vec	0.84	0.79	0.78	0.79
BERT	**0.87**	**0.84**	**0.84**	**0.84**
CountVectorizer	0.83	0.78	0.78	0.78

The evaluation results are presented in Table 4.1. Based on the result Table 4.1, We observe that the BERT model has slightly outperformed the other word embedding model. Its performance is Precision: 0.87, Recall: 0.84, F1 score: 0.84, and Accuracy:0.84 which is the highest performance compared to other word embedding methods.

Table 4.2: Naive Bayes Evaluation result table of word embedding methods.

Word Embedding Model	**Precision**	**Recall**	**F1**	**Accuracy**
TF-IDF	0.89	0.88	0.88	0.88
Word2Vec	0.99	1.00	0.99	1.00
BERT	**1.00**	**1.00**	**1.00**	**1.00**
CountVectorizer	0.71	0.58	0.59	0.58

Table 4.2 shows the evaluation result of the Naive Bayes. We observe that the BERT model has perfect performance in which precision, recall, f1 score, and accuracy are 1.00. We have confirmed that it is not biased since other methods have different results. Besides, Word2Vec also has a very good performance which is near to 1.00 in Precision, and F1 score.

Table 4.3: KNN Evaluation result table of word embedding methods.

Word Embedding Model	**Precision**	**Recall**	**F1**	**Accuracy**
TF-IDF	0.69	0.65	0.66	0.65
Word2Vec	0.71	0.70	0.69	0.70
BERT	**0.73**	**0.72**	**0.72**	**0.72**
CountVectorizer	0.68	0.65	0.66	0.65

Table 4.3 shows the evaluation result of the KNN. In this classification method, all the word embedding methods have a slightly low performance compared to the Random Forest, and Naive Bayes, but BERT still has the highest performance which is Precision: 0.73, Recall: 0.72, F1 score: 0.72, and Accuracy: 0.72.

In conclusion, BERT has the highest performance compared to TF-IDF, Word2Vec, and CountVectorizer based on the 3 classification models. Based on our observation, BERT will outperform other methods because it is a pre-trained model which is trained on 2.5 billion words and it uses bi-directional learning to predict the words from both left to right and right to left context. Besides, Next Sentences Prediction (NSP) training helps the model to understand the relationship between sentences.

4.6 Food Sentiment of Review

After implementing the Sentiment Analysis and Named Entity Recognition in the review, we found that only more than 2000 restaurants have food with good sentiment, and about 400 restaurants have food with bad sentiment. Based on the situation above, we can conclude that good sentiment is more than bad sentiment in this dataset. Although the number of restaurants has food with sentiment is less, the food found in the restaurant is accurate.

	name&num	FOOD
0	Tall Blonde French, 173	'chocolate shake'
2	Chung Wah, 179	'dragon prawn'
3	Stoner, 205	'cheesy burger', 'okish'
10	Stoner, 297	'cheesy burger', 'okish'
13	Ice Cream Works, 308	'ice cream'
...	...	...
2764	The Barn - Bar & Kitchen, 13761	'frnds'
2765	Stoner, 13801	'flavors'
2766	Anand Sweets And Savouries, 13837	'sugar cravings'
2767	Stoned Monkey, 13896	'icecreams'
2768	Pizza Hut, 13910	'toppings'

Figure 4.25: The example restaurant with good food.

	name&num	FOOD
0	Hammered, 303	'peri peri chicken wings', 'Drinks'
4	Hammered, 440	'peri peri chicken wings', 'Drinks'
9	Hammered, 738	'peri peri chicken wings', 'Drinks'
15	Hammered, 1161	'peri peri chicken wings', 'Drinks'
21	Hammered, 1746	'peri peri chicken wings', 'Drinks'
...	...	...
324	MISU, 8578	'dumplings'
332	MISU, 9812	'dumplings'
340	MISU, 12282	'dumplings'
344	MISU, 12520	'dumplings'
351	MISU, 12623	'dumplings'

Figure 4.26: The example restaurant with bad food.

4.6 Final Result of Restaurant Recommendation System

In Figure 4.27, the dataset with good food and bad food which is found using the NER and POS is combined with the BERT model where the BERT model has the best performance among the word embedding methods. While the restaurant does not have any good or bad food, the system will mention that there are no good food and bad food in the system like the “Recommended” and “Not Recommended” columns. Else, the food with good or bad sentiments is presented in the “Recommended” and “Not Recommended” columns.

cuisines	Restaurant_Name	Recommended	Not Recommended
North Indian, Chinese	The Royal Corner - Pai Viceroy	'Cheese garlic balls'	-
Arabian, Fast Food	The Shawarma Inc	-	-
North Indian	Riwaz - The Ritz-Carlton	-	-
North Indian	Savitha Family Restaurant	-	-
North Indian, Mughlai	The Spice Bazaar	-	-
North Indian, Chinese, Biryani, Rolls	Garma Garam	'lip smackingly delicious'	-
Desserts, Beverages	Smoor	-	-
North Indian	Oye Amritsar	-	-
North Indian, South Indian, Seafood, Biryani, Chinese	Aramane Restaurant	-	-
Cafe, Desserts	Waffle Magic	-	-

Figure 4.27: The recommendation model defining the sentiment of the food on ‘North Indian, Chinese’ cuisines.

cuisines	Restaurant_Name	Recommended	Not Recommended
North Indian, Street Food, Mithai	Laddoos	'samosas', 'kachoiri'	-
Biryani, Hyderabadi, Chinese	Biryani Palace	-	-
Biryani, Andhra, North Indian, Chinese	Shanmukha	-	-
North Indian	Kataria's Pakwan	-	-
Asian, Chinese, Continental, Italian	Thyme & Whisk	'broccoli', 'almond soup'	-
American, North Indian, Chinese	The Barn - Bar & Kitchen	'frnds'	-
Bakery, Desserts, Beverages	Mad Over Donuts	-	-
Continental, Chinese, Fast Food	Amber Rush Restobar	-	-
Cafe, Chinese, Pizza, North Indian, Burger	Entropy Cafe	-	-
Cafe, Beverages, Healthy Food, Juices	Vitamin Palace	-	-

Figure 4.28: The recommendation model defining the sentiment of the food on ‘North Indian, Street Food, Mithai’ cuisines.

Chapter 5 Conclusion

In this work, we have proposed a content-based restaurant recommendation system using Natural Language Processing (NLP). Word Embedding methods: TF-IDF, BERT, Word2Vec, and Countvectorizer are used which is one of the subfields of the NLP to train the recommendation system to find similar restaurants using Cosine Similarity through the reviews list of the restaurants from the Bangalore Restaurant Dataset. To evaluate the word embedding model, the supervised proxy problem is implemented by training three classification models: Random Forest, Naive Bayes, and KNN. Based on the evaluation result from these 3 classification models, BERT has outperformed other word embedding models in terms of Accuracy, Precision, Recall, and F1 scores. Especially in the Naive Bayes classification model, it performs perfectly with 100% of Accuracy, Precision, Recall, and F1 scores. To enhance the recommendation system, we implemented sentiment analysis and named entity recognition to find out the good or bad food entity from the dataset. We had used a pre-trained model from DeepNote.com to define the food entity from dataset. During the sentiment analysis process, there are 5 types of parts of speech from positive and negative reviews with food entities that have been found from the sampling of the dataset. Lastly, these parts of speech are used to define the sentiment of the food entity.

5.1 Limitation and Challenges

The limitation of this work is the sentiment analysis. During the sentiment analysis process, we used the Part of Speech to define the sentiment of the food review, but we cannot define the sentiment perfectly while the polarity of the review is represented with multiple sentences. In this case, we will continue to enhance the system in the Sentiment Analysis field in the future. Next, the challenge of this work is the dataset used in this work. There is no labeled dataset as the ground truth to test the training dataset. So, the evaluation is the most challenging part of this work.

References

[1] Ananda Babu, J., Vinay, D. R., Kumaraswamy, B. v, Chandra, C., & Basavaraddi, S. (2021). To cite this article: J Ananda babu et al 2021. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1964/4/042081

[2]Alexander Geiger. (2021). *Evaluating recommender systems in absence of labeled data | SAP Blogs*. (n.d.). Retrieved October 12, 2021, from https://blogs.sap.com/2021/03/29/evaluating-recommender-systems-in-absence-of-labeled-data/

[3] Asiful Huda, S. M., Shoikot, M. M., Hossain, M. A., & Ila, I. J. (2019). An effective machine learning approach for sentiment analysis on popular restaurant reviews in Bangladesh. *Proceedings - 2019 1st International Conference on Artificial Intelligence and Data Sciences, AiDAS 2019*, 170–173. https://doi.org/10.1109/AIDAS47888.2019.8970976

[4] Badriyah, T., Azvy, S., Yuwono, W., & Syarif, I. (2018). Recommendation system for property search using content based filtering method. *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, *2018-January*, 25–29. https://doi.org/10.1109/ICOIACT.2018.8350801

[5] Bagher, R. C., Hassanpour, H., & Mashayekhi, H. (2017). User trends modeling for a content-based recommender system. *Expert Systems with Applications*, *87*, 209–219. https://doi.org/10.1016/J.ESWA.2017.06.020

[6] Cai, X., Hu, Z., Zhao, P., Zhang, W. S., & Chen, J. (2020). A hybrid recommendation system with many-objective evolutionary algorithm. *Expert Systems with Applications*, *159*, 113648. https://doi.org/10.1016/J.ESWA.2020.113648

[7] Chang, H., Wu, W., Shu, X., Dong, Z., Liu, Z., & Weng, J. (2021). Application of Content Based Recommendation System in Homestay. *Proceedings - 2021 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2021*, 493–496. https://doi.org/10.1109/ICITBS53129.2021.00126

[8] Chen Dong Liu, X., Lei, C., Li Zheng-Jun Zha Zhiwei Xiong, R., Chen, X., Liu, D., Li, R., & Zha, Z.-J. (n.d.). BERT4SessRec: Content-Based Video Relevance Prediction with Bidirectional Encoder Representations from Transformer Bidirectional Encoder Representations from Transformer (BERT); deep learning; session-based recommendation ACM Reference Format. *Proceedings of the 27th ACM International Conference on Multimedia*. https://doi.org/10.1145/3343031

[9] Chu, W.-T., & Tsai, Y.-L. (2017). A hybrid recommendation system considering visual information for predicting favorite restaurants. *World Wide Web 2017 20:6*, *20*(6), 1313–1331. https://doi.org/10.1007/S11280-017-0437-1

[10] Cristy Patty, J., Thea Kirana, E., Sandra Diamond Khrismayanti Giri, M., Teknik Informatika, M., & Atma Jaya Yogyakarta, U. (2018). Recommendations System for Purchase of Cosmetics Using Content-Based Filtering. *International Journal of Computer Engineering and Information Technology*, *10*(1), 1–5. www.google.com

[11] del Carmenrodriguez-Hernandez, M., Del-Hoyo-Alonso, R., Ilarri, S., Montanes-Salas, R. M., & Sabroso-Lasa, S. (2020). An Experimental Evaluation of Content-based Recommendation Systems: Can Linked Data and BERT Help? *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2020-November*. https://doi.org/10.1109/AICCSA50499.2020.9316466

[12] Gandhi, M., Gandhi, M., & Gandhi, S. (2019). An Enhanced Approach for Tourism Recommendation System using Hybrid Filtering and Association Rule Mining. *Asian Journal For Convergence In Technology (AJCT) ISSN -2350-1146*, *0*(0). https://asianssr.org/index.php/ajct/article/view/753

[13] Gomathi, R. M., Ajitha, P., Krishna, G. H. S., & Pranay, I. H. (2019). Restaurant recommendation system for user preference and services based on rating and amenities. *ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings*. https://doi.org/10.1109/ICCIDS.2019.8862048

[14] Hassan, A. K. A., & Abdulwahhab, A. B. A. (2017). Reviews Sentiment analysis for collaborative recommender system. *Kurdistan Journal of Applied Research*, *2*(3), 87–91. https://doi.org/10.24017/SCIENCE.2017.3.22

[15] Jalan, K., & Gawande, K. (2018). Context-aware hotel recommendation system based on hybrid approach to mitigate cold-start-problem. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*, 2364–2370. https://doi.org/10.1109/ICECDS.2017.8389875

[16] Jelodar, H., Wang, Y., Rabbani, M., & Ayobi, S. (2019). *Natural Language Processing via LDA Topic Model in Recommendation Systems*.

[17] Kapoor, N., Vishal, S., & K. S., K. (2020). *Movie Recommendation System Using NLP Tools*. 883–888. https://doi.org/10.1109/ICCES48766.2020.9137993

[18] Kaviani, M., & Rahmani, H. (2020). EmHash: Hashtag Recommendation using Neural Network based on BERT Embedding. *2020 6th International Conference on Web Research, ICWR 2020*, 113–118. https://doi.org/10.1109/ICWR49608.2020.9122275

[19] Koetphrom, N., Charusangvittaya, P., & Sutivong, D. (2018). Comparing Filtering Techniques in Restaurant Recommendation System. *2018 2nd International Conference on Engineering Innovation, ICEI 2018*, 46–51. https://doi.org/10.1109/ICEI18.2018.8448528

[20] Mishra, R. K., Urolagin, S., & Jothi, A. A. J. (2019). A Sentiment analysis-based hotel recommendation using TF-IDF Approach. *Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019*, 811–815. https://doi.org/10.1109/ICCIKE47802.2019.9004385

[21] Narwani, B., Nawani, J., Kejriwal, S., Shankarmani, R., & Patel, S. (2020). in A Hybrid Recommendation System for Restaurants A HYBRID RECOMMENDATION SYSTEM FOR RESTAURANTS. *International Journal of Advances in Electronics and Computer Science*, *7*, 2394–2835. http://iraj.

[22] Patidar, C. P., Sharma, M., & Katara, Y. (2020). Hybrid News Recommendation System using TF-IDF and Machine Learning Approach. *International Journal of Innovative Science and Research Technology*, *5*(10). www.ijisrt.com

[23] Patra, B. G., Maroufy, V., Soltanalizadeh, B., Deng, N., Zheng, W. J., Roberts, K., & Wu, H. (2020). A content-based literature recommendation system for datasets to improve data reusability – A case study on Gene Expression Omnibus (GEO) datasets. *Journal of Biomedical Informatics*, *104*, 103399. https://doi.org/10.1016/J.JBI.2020.103399

[24] Pradeep, N., Rao Mangalore, K. K., Rajpal, B., Prasad, N., & Shastri, R. (2020). Content based movie recommendation system. *International Journal of Research in Industrial Engineering*, *9*(4), 337–348. https://doi.org/10.22105/RIEJ.2020.259302.1156

[25] Rutkowski, T., Romanowski, J., Woldan, P., Staszewski, P., Nielek, R., & Rutkowski, L. (2018). A content-based recommendation system using neuro-fuzzy approach. *IEEE International Conference on Fuzzy Systems*, *2018-July*. https://doi.org/10.1109/FUZZ-IEEE.2018.8491543

[26] Shrirame, V., Sabade, J., Soneta, H., & Vijayalakshmi, M. (2020). Consumer Behavior Analytics using Machine Learning Algorithms. *Proceedings of CONECCT 2020 - 6th IEEE International Conference on Electronics, Computing and Communication Technologies*. https://doi.org/10.1109/CONECCT50063.2020.9198562

[27] Singhal, A., Sinha, P., & Pant, R. (2017). Use of Deep Learning in Modern Recommendation System: A Summary of Recent Works. *International Journal of Computer Applications*, *180*(7), 17–22. https://doi.org/10.5120/ijca2017916055

[28] Sunandana, G., Reshma, M., Pratyusha, Y., Kommineni, M., & Gogulamudi, S. (2021). Movie recommendation system using enhanced content-based filtering algorithm based on user demographic data. *Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021*. https://doi.org/10.1109/ICCES51350.2021.9489125

[29] Vuong Nguyen, L., Nguyen, T.-H., & Jung, J. J. (n.d.). Content-Based Collaborative Filtering using Word Embedding: A Case Study on Movie Recommendation. *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*. https://doi.org/10.1145/3400286

[30] Wang, T., & Fu, Y. (2020). Item-based Collaborative Filtering with BERT. *Online*, 54–58. https://doi.org/10.18653/V1/2020.ECNLP-1.8

[31] Yoon, Y. C., & Lee, J. W. (2018). Movie Recommendation Using Metadata Based Word2Vec Algorithm. *2018 International Conference on Platform Technology and Service, PlatCon 2018*. https://doi.org/10.1109/PLATCON.2018.8472729