Improving Customer Service Efficiency Using Generative Artificial Intelligence

Wei Fan

Master's Thesis

MEng in Big Data Analytics

2024

Master’s Thesis

Wei Fan

Improving Customer Service Efficiency Using Generative Artificial Intelligence.

Arcada University of Applied Sciences: MEng in Big Data Analytics, 2024.

Commissioned by:

N/A

Abstract:

Social media has presented customer relationship management with a challenge while providing companies with a tool to maintain their relationships with customers. Effort and time in processing customers’ feedback may increase as the widespread use of smart mobile devices. The evolution of generative artificial intelligence (generative AI) is a possible solution to improve the efficiency of customer service. However, existing discussions mainly focus on how chat tools can benefit from generative AI. There is a lack of exploration of how generative AI can assist companies provide managerial responses for customers, which is an important part in customer relationship management and has different requirements from chat tools. This study examines the capability of generative AI in understanding customer reviews and generating managerial responses automatically by fine-tuning two language models. A dataset of customer reviews of restaurants is used. Results indicate that even with a relatively small model, generated responses are still satisfied. The current study also compares different methods to predict whether customers’ reviews would obtain managerial responses. It is observed that the patterns of making decisions about whether to reply to customer reviews differ across restaurants.

Keywords: Online Reviews, Managerial Responses, Generative AI, Large Language Models

Contents

1 Introduction

In recent decades, the Internet and the widespread use of smartphones enable consumers to share their experience about products or services online. They can write online reviews to make recommendations or express dissatisfaction. These online reviews help potential consumers obtain product information efficiently. They also help consumers make decisions and plan their purchases because the quality of services is often difficult to observe before consumption. Moreover, online reviews can affect the sales of products (S. Liu et al., 2013). It has been concluded that online review is the key factor of 20%-50% of purchases (Mathwick & Mosteller, 2017). Online reviews also help practitioners realize customers' demand and address issues that are typically escalated to the management (Schuckert et al., 2015), thereby allowing them to improve their products and services efficiently. By mining online reviews, practitioners can understand customers' behaviors and preferences, from which they can construct proper marketing strategies.

Moreover, social media websites allow customers to communicate with product or service providers easily. Besides the traditional ways like telephone and email, people can contact brands on social media. There are companies in different industries operating official accounts on social media for publishing news and marketing. Customers can raise queries and seek problem solutions by contacting the accounts of these companies. Staff in charge of these accounts can respond to customers' inquiries and handle complaints online. Some booking or rating websites also enable product or service providers to reply to customer reviews. These online managerial responses improve the satisfaction of customers and the efficiency of customer service. Potential customers' purchase intention may be influenced because these responses are usually publicly available. Managerial responses are expected to enhance an enterprise's relationship with satisfied customers, but also to reduce the possible damage of negative customer reviews. Typically, companies selectively respond to customer reviews according to review contents and their resources in terms of labor and time.

Industry practitioners keep seeking ways to improve their efficiency to manage customer relationships. For example, many companies have enabled chatbots on their websites for answering customers' basic inquiries. In recent years, natural language processing (NLP) technologies have helped companies in customer service. With NLP, companies can comprehend the feedback of their products or service from their customers efficiently. Moreover, responses to customers' inquiries can be automatically generated (Olujimi & Ade-Ibijola, 2023), which can be used as references to help improve the quality and speed of the human. In 2022, ChatGPT made a splash when it was released. Its success makes people realize the capabilities of generative AI to generate natural, coherent, and contextually relevant conversations, which may provide a better solution for the interactions with customers. In addition, Generative AI chatbots can analyze customer data to understand the context of each customer inquiry and use natural language processing techniques to craft a personalized response.

Building upon the successes and learnings from ChatGPT, researchers have conducted investigations of enhancing the capabilities of generative AI. There have been efforts in discussing the possible applications of generative AI in customer service, such as identification of trends and chatbots (Korzynski et al., 2023; Verma & Kumari, 2023). However, there is little research focused on the practical use of generative AI in customer service, especially the managerial response to customers. The previous discussions about linguistic features of managerial response and chatbot indicate that their requirements are different (M. Li & Wang, 2023; Xu & Zhao, 2022). There is also a gap in comparing actual managerial responses with generated text and investigating the efficiency of such practical application. Therefore, this research aims to get insight into how enterprises can properly grasp the opportunities offered by generative AI. A dataset which contains online reviews of customers, managerial responses, and other attributes will be used. The objectives of thesis are as follows:

1. To explore the aspects of a service that consumers care about in online reviews.
2. To investigate the possibility to classify the customer reviews using machine learning and decide which ones would be replied to.
3. To fine-tune a large language model to generate managerial responses for customer reviews.

This thesis is organized as follows. In preliminary and related works, we review the literature on the study of customer relationship management. Then the methodology used in this thesis is presented. The results for topic modeling, classification of customer reviews, and the fine-tuned model are provided, followed by the discussion of the application of generative AI in customer service. Lastly, we conclude this study and discuss future research.

2 Preliminary and related works

2.1 Online reviews of customers

As social networking sites emerge, which allow people to share their ideas online, communications over the internet have become an important way to acquire information of products and services. Such communication is defined as electronic word-of-mouth (eWOM) (Litvin et al., 2008). It is summarized that the eWOM communications include the ones between consumers and the ones between consumers and producers. The communication channels consist of Emails, Blogs, rating websites, virtual communities or groups, and instant chat tools. Serra Cantallops and Salvi (2014) conclude the factors in generating reviews of hotels, which include customer satisfaction/dissatisfaction, service quality, helping other vacationers, social identity, and so on. On review websites users can make recommendations and express complaints about specific products or services. These online reviews provide people with information to make decisions. Works have been done to mine online reviews to explore consumers' sentiments, satisfaction levels, expectations, and so on. Xie, Chen, et al. (2016) have demonstrated the factors of reviews which influence offline hotel popularity using regression in their work. Jiménez and Mendoza (2013) have investigated the influence of online reviews and found that consumers assess the credibility of reviews with different criterions for different categories of products. Yu et al. (2010) have extracted sentiment and quality factors from online reviews to predict sales of movies.

Researchers have also examined the influence of eWOM on consumers' purchase behavior. Vermeulen and Seegers (2009) have demonstrated that both positive and negative reviews can increase customers' awareness of hotels. In the paper of Kudeshia and Kumar (2017), it is found that positive user generated contents on social media have significant influence on consumers' attitude to brands. In addition, Sparks and Browning (2011) have found that positive eWOM information can increase people's trust and hotel booking intention. Tran's (2020) study has also confirmed online reviews' positive effect on purchase intention of hotel reservation. Qiu and Zhang (2023) have reviewed research articles and identified the factors of online reviews which can influence customers' purchase intention. The factors include review-related factors, such as review rating, review volume, and review credibility, and reviewer-related factors, i.e. source trustworthiness. Moreover, culture, product types and review platform type can affect the influence of online reviews. It is suggested that customer reviews are more valuable for experience products.

For product and service providers, mining online reviews enables them to understand their customers' preferences and requirements. Guo et al. (2017) used latent dirichlet allocation (LDA) to extract topics from hotel reviews and find out important aspects which influence customers' satisfaction with machine learning methods. Jia (2020) has compared the satisfaction of two groups of customers by mining their reviews on social media websites with LDA. Product and service providers can also make improvements based on the outcomes of online reviews mining. Qi et al. (2016) have proposed a method to identify reviews helpful for developing new products in their work. M. Zhang et al. (2021) focused on the expectations of customers and utilized deep learning to mine innovative ideas from online reviews. Such ideas can be helpful in improving and developing products.

2.2 Managerial responses

Social media websites on which users can write online reviews may also allow companies to reply to their customers. Through this channel, companies can express their appreciation or deal with customers' complaints. Also, it is a way to interact with customers, which can strengthen their relationship to customers and help brand reputation (Van Noort & Willemsen, 2012) . By constructing a regression model and using data from TripAdvisor, researchers have demonstrated that managerial response function use increases hotel ratings. Moreover, the subsequent review quantity rises by 17.3%, which may indicate the increment of customers (Xie, Zhang, et al., 2016). However, industry practitioners would selectively reply to the reviews they received. The reason may be that managers do not think it is necessary to respond to some reviews based on their contents. Another possible reason is the limited time or labor. It has been found that about 19% of reviews obtain managerial responses in a dataset of hotel reviews (Xie, Zhang, et al., 2016). It is also mentioned that the percentage of reviews obtained managerial responses has increased rapidly, and the ratio of their work is around 49% (Wu et al., 2023). Researchers have explored the factors which influence companies' decisions to respond to specific reviews. In the work of X. Liu and Law (2019), it is noticed that managers would prioritize influential and negative reviews and pay more effort to respond to such reviews. Qiu and Zhang (2023) also suggest companies to respond to negative review promptly to reduce negative impact. Wu et al. (2023) have demonstrated that reviews related to changeable attributes and covering more topics are more likely to receive responses. Thus, managers need to make efforts to prioritize or classify reviews, in addition to responses writing.

Researchers have discussed the linguistic characteristics of managerial responses. It has been found that sentiment, subjectivity, diversity and other features of managerial responses have influence on customers' satisfaction (Xu & Zhao, 2022). Deng and Ravichandran (2023) suggest that for different types of customer reviews, tailored or template responses can be applied. By calculating language matching scores for a dataset of hotel reviews, Ren et al. (2024) have shown that language style matching between review and managerial response has a positive effect on subsequent ratings. Therefore, it is important for practitioners to consider the language style when replying to customer reviews, not just using simple templates. This requirement would increase the workload of managers.

2.3 Generative AI

The development of artificial intelligence has helped the improvement of various fields, such as supply chain management (Pournader et al., 2021), healthcare industry (Hee Lee & Yoon, 2021), and hospitality industry (Prentice et al., 2020). AI can be used in prediction analysis, fraud detection, computer vision and so on. One of the usages of AI is language understanding, include translation, question answering, and text summarization. With the release of ChatGPT, the public get to realize the potential usages of generative AI, like text, image, and audio generation. Researchers have discussed the applications of generative AI in different fields. Rodler et al. (2024) argue that generative AI can be used for surgical planning by extracting patient information, literature, and guidelines. Due to the increasing amount and complexity of medical data, efficiency can be improved with generative AI compared to traditional methods. Aguinis et al. (2024) provide guidelines to create prompts in order to generate valuable recommendations for assisting tasks of human resource management. In the work of Korzynski et al. (2023) it is mentioned that generative AI would affect managerial work, including automating customer service interactions. Consider the challenges discussed in the above section, it is possible that customers service efficiency would be improved with the help of generative AI.

Generative AI is able to understand the context of customers' inquiries and create personalized responses (Dogru et al., 2023). There have been works on the use of generative AI in customer service. Castillo et al. (2021) have explored the failure of AI-powered chatbot experience by interviewing customers. They concluded five reasons of failed interaction, such as limited functionality of chatbots and the lack of alternative human support. Olujimi and Ade-Ibijola (2023) have reviewed papers on the applications of NLP technique in customer service. They concluded that chatbot is the broadest used application, which can be trained to answer customers' queries. With the implementation of AI-powered chatbots, manual efforts can be minimized, while the inaccurate response is a challenge. In the study of Flavián et al. (2022), a survey is conducted to investigate people's intention to use AI-powered investment advisor. They have found that customers with more technology discomfort may show more tendency to use this service. Prentice et al. (2020) worked on the questionnaire data of customers and concluded that the service quality of AI-powered chatbot would influence customer trust and loyalty.

There are also works on building or improving tools for customer service. In the study of C. Liu et al. (2020), it is noticed that if a customer expresses a request in more than one turn, chatbots providing instant replies at every turn may misunderstand and generate inaccurate responses. A response trigger model is proposed to decide the timing that an intelligent chatbot generates responses. Wang et al. (2021) pretrained a model based on a BERT model with chat dialogues between customers and customer service agents. It is found that a domain-specific model can improve the performance in specific tasks. S. Liu et al. (2023) have proposed two methods to optimize pretrained language models for the applications in customer service. Their models are evaluated with eight datasets covering different tasks, such as sentiment analysis and query matching.

Existing studies focus on utilizing AI-powered chatbots or query answering to improve customer service. There is little works on assisting managerial responses with generative AI. The possibility of this application has been discussed by Dogru et al. (2023). It is suggested that generative AI can identify the sentiments in online reviews and provide real-time responses. It is also mentioned that companies should be careful when using the contents from generative AI to avoid the destruction of trust. However, there is still a lack of practice in generating managerial responses with generative AI. The effectiveness and feasibility analysis of such an application of generative AI can be provided through experiments on this practice. This is the problem that this thesis would like to solve.

2.4 Summary

Previous research has demonstrated that online customer reviews can influence the sales of products or services, and managerial responses from industry practitioners help maintain and strengthen customer relationships. Monitoring and replying to customer reviews are important tasks in customer service. However, reading reviews and writing responses are time and labor consuming. As the development of generative AI, customer service may benefit from domain-specific language models. Large language models can understand the context of customer reviews and generate conversational and personalized text, which meet the requirements of a good managerial response. Nevertheless, existing research mainly focuses on chatbots or customer conversations. There is a gap in generating managerial response with AI. Moreover, a lack exists in the experiments of generative AI in customer service using real data. This study will provide such an attempt and discuss the feasibility of using generative AI in assisting customer relationship management on online customer reviews.

3 Methodology

3.1 Dataset

This work will use the data published by Yan et al. (2023) and J. Li et al. (2022). This dataset contains customer reviews and metadata for businesses in United States from Google Maps up to September 2021. Google Maps is a web mapping service, available on both desktop computers and mobile devices. It is widely used for navigation and exploring places. Users can rate and leave reviews for businesses listed on Google Maps. Google Maps also provides businesses with the opportunity to update their information, e.g., opening hours and contact details. Moreover, they can respond to customer reviews. Figure 1 is the profile of a restaurant on Google Maps. The dataset also includes responses for these reviews from these businesses, if any. The basic information of the business is recorded in the dataset, as well.

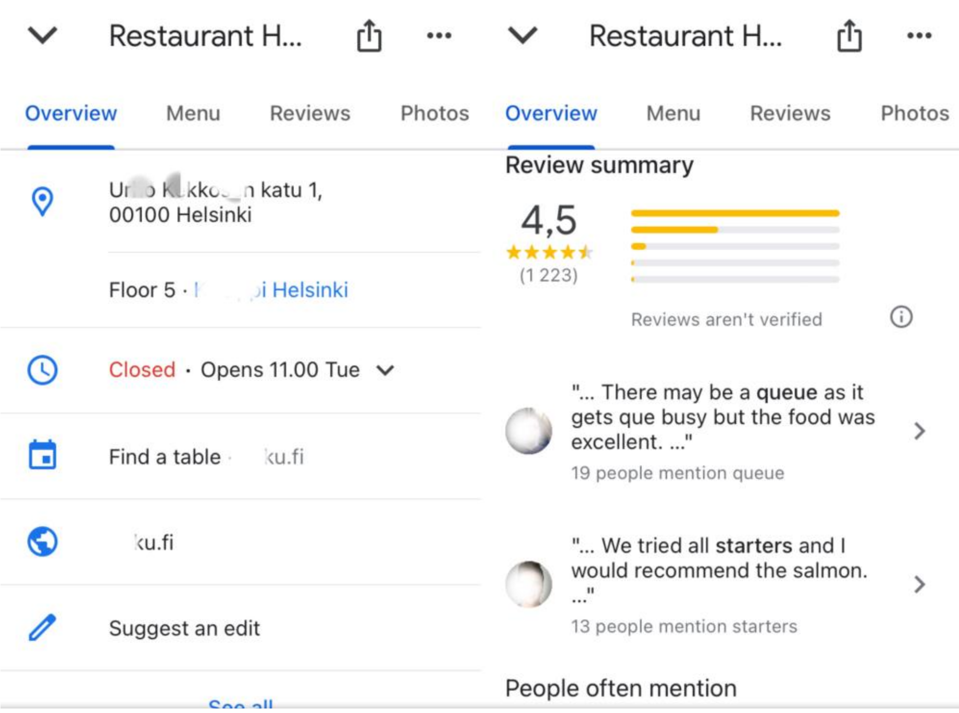


Figure 1. An example of the businesses on Google Maps.

In this thesis, we use the data of restaurants in Hawaii, which covers 4286 restaurants and contains 1,135,436 reviews. The data of a specific category of business can be obtained by filtering out the businesses that contain the string "restaurant" in their category information. For example, "Restaurant," "Seafood restaurant," and "Pizza restaurant" are included in the dataset. Other food and drink related businesses, e.g., cafes and bars, are not in our dataset because they do not have the word "restaurant" in their category information. The reviews for restaurants cover about 36% of the reviews for all businesses in Hawaii.

After a statistical analysis of the dataset, three tasks will follow according to the three research questions of this thesis. The customer reviews for all restaurants will be used in the topic modelling task to understand customers' opinions. This study mainly focuses on managerial responses to customer reviews, so we also extract two subsets for solving different problems, as shown in Figure 2. In subset A, every restaurant has replied to at least one review. Subset A will be used to model how managers prioritize the reviews for responding, because only the restaurants with response history have the prioritizing process. This subset contains 551,139 reviews for 1367 restaurants. 106, 094 of the reviews have received managerial responses from the restaurants. Subset B is made of these review-response pairs and the corresponding attributes of customers and restaurants from subset A, and it will be used in the task of fine-tuning large language models.

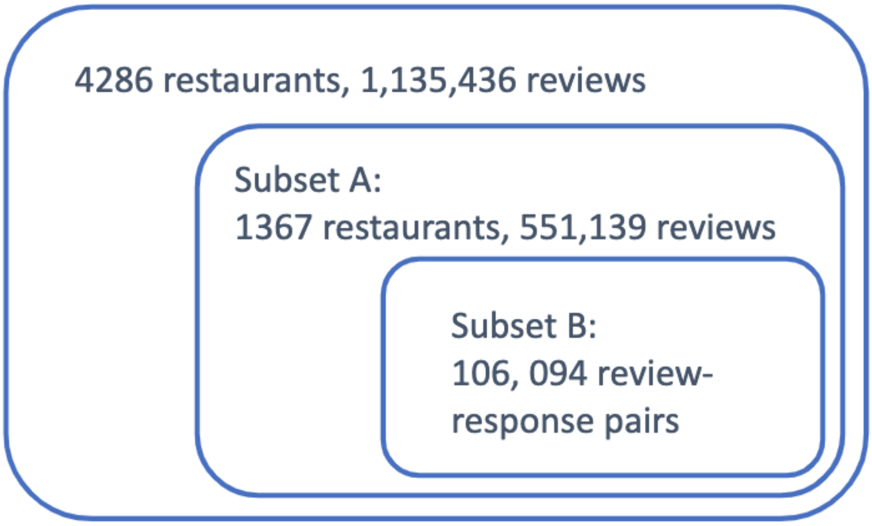


Figure 2. The dataset and two subsets used in this study.

Some information is dropped from the dataset, such as the uniform resource locators (URL) of the pictures uploaded as part of reviews, price level, and the latitude and longitude of the restaurant. Three columns of bool-value data are added to record the features of reviews. For example, we drop the URLs of pictures, but add the status of whether a review contains pictures. Such features will be used in the statistical analysis and the task of review classification. Table 1 lists the information used in this study.

Table 1. Information in the dataset

Information of customers	User ID Reviewer name
Information of restaurants	Google Maps ID Restaurant name Average rating Number of reviews
Information of reviews	Rating With review text or not (1 or 0) Review Text, if any With pictures or not (1 or 0) Responded or not (1 or 0) Response text, if any

3.2 Topic modelling

To answer the first research question of this study, which aims to explore what aspects customers would care about, we will use latent dirichlet allocation (LDA) to discover the topics in customer reviews of the whole restaurant dataset. Introduced by Blei et al.(2003), LDA is a popular topic modelling technique to find the topics and the corresponding words in the identified topics. There are other topic modelling methods, such as LSA (latent semantic analysis), CTM (correlated topic model), and some methods using neural networks. However, the results of LDA are easier to interpret and LDA requires less computational resources (Yu & Xiang, 2023). Although LDA may have difficulty in capturing complex topic distributions as effectively as neural networks, it is selected for this task because of its interpretability and simplicity while the size of our dataset is large. LDA assumes that each document is a mixture of latent topics, and each topic can be represented by a distribution of words. With a given number of topics, it first assigns a random topic for each word in each document. Then it iteratively updates the topic assignments based on topic and word distributions of the model and the word frequencies in documents. LDA can be used to understand review text and provide information on customers’ experiences from the dataset. In this thesis, the LdaModel class in the python library Gensim is used.

Before applying LDA, the online reviews need to be preprocessed. In the dataset of Google Maps reviews, some reviews were written in languages other than English. These reviews include both the translated text and the original reviews, in a format of “(Translated by Google) <English review> (Original) <review in other languages>.” We keep only the translated reviews to avoid “Translated by Google” and “Original” being considered as topics. Then the documents need to be cleaned. The steps include removal of English stop words, clearing punctuation, and lemmatizing the word in documents. After that, a dictionary of the bag of words and a term-document matrix are created as the inputs of LDA model.

To apply LDA, it is necessary to select the number of topics as a parameter of the topic model. However, there is not a public accepted method to determine the optimal number of topics without prior knowledge of the data (Gan & Qi, 2021). The metrics to choose the number of topics is not the focus of this study, and we will not discuss this in this thesis. To decide the value of this parameter, we will start from 2 and observe the distance map of the detected topics. Then we will do an iteration by increasing the value by 1 and regenerating the distance map. If the detected topics overlap and the keywords for each topic repeat a lot, which means a poor performance of topic modelling, we will stop the increment and choose the optimal number of topics based on the observations.

Table 2. Possible words in different aspects of restaurant reviews

Aspects			
Food	Service	Atmosphere	Value
• Food item • Taste • Ingredient • Drink	• Staff • Order • Time	• Bathroom • Furniture • Music • View	• Price • Expensive • Worth

In previous works, food, value for money (price), atmosphere, and service are the frequently studied features of restaurants that influence consumers’ selection and experience (C. H. S. Liu et al., 2014; Rhee et al., 2016). Therefore, it is assumed that these four aspects should be found in the detected topics of the dataset of restaurant reviews.

Considering that customers can write reviews in a free style on Google Maps, and they do not need to strictly follow a framework, multiple aspects are possible to appear in a single topic. According to previous research of Panchendrarajan et al. (2016) and Abdullah et al. (2021), some example words in each aspect are listed in Table 2.

The results of topic modelling will provide the probability distribution of words in each topic. From these results, the experience of customers and the aspects of restaurants they care about can be understood.

3.3 Classification of reviews

As it is discussed in the literature review section, managers usually prioritize online reviews to decide which are more necessary to respond to. However, both prioritizing reviews and writing responses need time and labor. The objective of review classification is to train a model to automatically filter out reviews of high priority based on their features. It can provide managers with references to decide whether to respond to a specific review, which would help improve efficiency. We exclude the restaurants without responses because they do not have this prioritizing process and there is no need to model their selecting behavior. In this study, subset A defined in Section 3.1 will be used for classification, i.e., the reviews of the restaurants with responding history. The label for classification is defined as whether a review is replied to by the restaurant. The labels of reviews without responses are 0 and others are 1. It is noticed that the dataset is imbalanced because the distribution of the label is uneven. The percentage of reviews with managerial responses is around 19%. To avoid building a model which is biased to classify majority class, i.e. reviews without responses, we downsample the majority class before model training.

There are numerical and text information for each review. Some managers may make reply decisions based on numerical information, such as ratings. Some managers may read the text before making decisions. This study will use different approaches to make classification and compare the accuracy. Table 3 lists the approaches using different techniques. Logistic regression and random forest models are two fundamental classification techniques. These two approaches will consider only numerical features. Logistic regression is an algorithm for binary classification tasks. Although it cannot capture non-linear relationships, it is easy to use, and the weights of the model can provide information about the influence of each feature. The probability that a review has a response (i.e. label is 1) can be determined by the sum of the products of the features and their corresponding weights. We will use the LogisticRegression class in scikit-learn library for model training, and it will find the optimal weights of features to minimize the cross-entropy loss between predictions and observations in training set. Three hyperparameters will be tuned with grid search to find a model with the best combination of hyperparameters.

Random forest is also a widely used classification method and it is a combination of multiple decision trees. Decision trees have tree structures and find important features in decision and their threshold by continuously dividing the multidimensional feature space into partitions. Random forest can reduce overfitting and achieve better performance by combining multiple decision trees and applying random sampling and random feature selection in training. In this thesis, the RandomForestClassifier class in scikit-learn library will be used. Number of trees, maximum depth of trees, and other two hyperparameters will be tuned to identify the model with the best result.

Table 3. Classification approaches

	Model	Used information
1	Logistic regression	• Rating of a review • With review text or not (1 or 0) • Picture attached or not (1 or 0) • Average rating of the restaurant • Number of reviews of the restaurant
2	Random forest	• Rating of a review • With review text or not (1 or 0) • Picture attached or not (1 or 0) • Average rating of the restaurant • Number of reviews of the restaurant
3	Neural network with numerical information and LSTM output of review text	• Rating of a review • With review text or not (1 or 0) • Picture attached or not (1 or 0) • Average rating of the restaurant • Number of reviews of the restaurant • Review text
4	Sequential classification using BERT	• Review text

Besides numerical features, managers may refer to the text comments in customer reviews and make responding decisions. The third approach will combine numerical features and the review text for classification. It is a neural network consisting of interconnected neurons. The architecture of this approach is explained in Figure 3. There are two input branches for this neural network. Numerical features can be used as inputs of one branch after normalization. Text of customer reviews are the input of another branch after tokenization and word embedding. For the reviews without text but only ratings, we fill their text value as “None”. Recurrent Neural Networks (RNNs) are commonly used to process text. Compared to traditional RNNs, Long Short-Term Memory (LSTM) networks can handle long-term dependencies and improve performance (Hochreiter & Schmidhuber, 1997). An LSTM layer is applied to the review text to capture semantic information. Then the outputs of two branches are merged as the inputs of an additional densely connected layer. This approach will be done with Keras library.

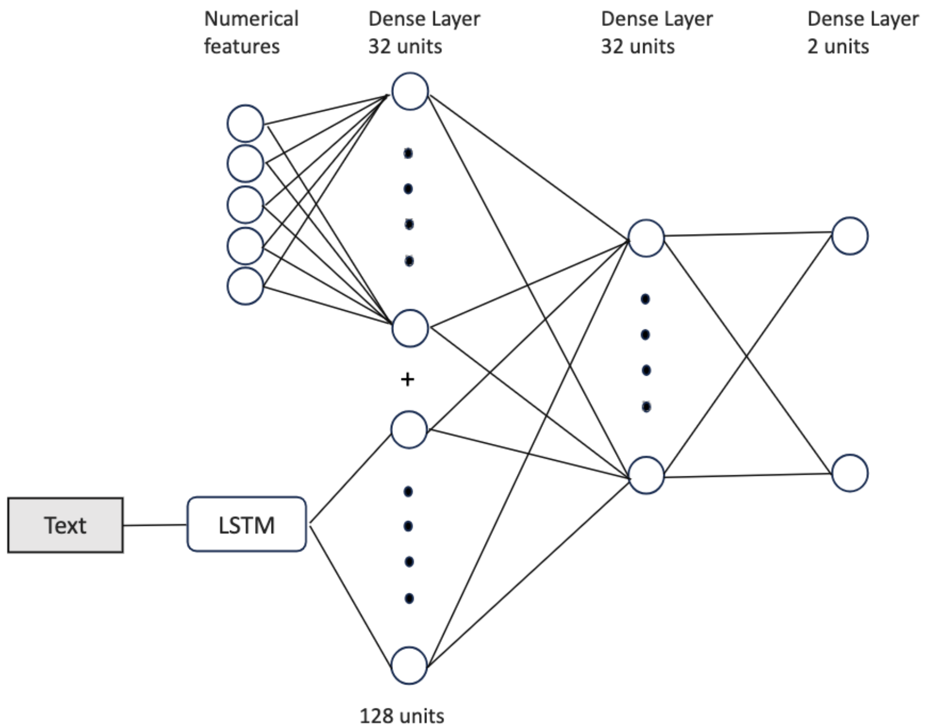


Figure 3. Architecture of the neural network using both numerical and text features for classification.

NLP technologies are helpful in extracting semantic information from texts and there have been various applications, such as scoring (Gutierrez-Bustamante & Espinosa-Leal, 2022). The last approach will use BERT (Bidirectional Encoder Representations from Transformers), a popular method in NLP, for tokenization and modeling. BERT is a pretrained language model developed by Devlin et al. (2018), which can acquire contextual meaning from text. It can be fine-tuned for classification task by adding an additional layer. BERT outperforms many other sequential classification models in various benchmark experiments (Piskorski et al., 2020). This approach will use only review text for classification. However, the dataset for this approach is large and would take a long time for training. Because of the limited resources, we will train the model with sampled data and the size is around 1% of the original size. The other three approaches will use the full subset A. The Transformers library in Python will be used for this approach. Accuracy will be used as the metric to measure the performance of the four approaches. Confusion matrices will also be provided for reference.

3.4 Generating responses with generative AI

The third problem this study wants to solve is to explore using generative AI to improve customer service efficiency. The first step is to prepare the context, which is the input text of large language models for response generating. Considering that the dataset used is composed of reviews and responses of different restaurants, the names of restaurants should be included in context. It is noticed that some restaurants would mention customers' names to strengthen their relationship with customers and provide personalized responses. Therefore, the names of customers (usernames on Google Maps) are also included in input context. The values of rating are also included because they can provide information about customers' overall perceived value. The review text, if any, is a part of the context.

There are many large language models for text generation. In this study, the pretrained Transformer based DialoGPT will be used. It is trained on conversations from online discussion platforms and provides interactive outputs in various domains (Y. Zhang et al., 2019). This model is selected for this task because of its small size. The time and resource cost to fine-tune this model are acceptable. Another reason is that this model is free. Different checkpoints of this model have been released by Microsoft and this study will fine-tune DialoGPT-small and medium. They have fewer parameters (117 million and 345 million) than DialoGPT-large (762 million) but are still able to generate coherent responses with less resources. For comparison, GPT 3, which is well known because of ChatGPT, has around 175 billion parameters, significantly more than DialoGPT.

Figure 4 shows the fine-tuning process of this project and Figure A.1 in appendices provides the pseudo-code of this process. This study uses review-response pairs in the dataset, and the pairs are split into train, validation, and test sets. The first two sets are used to fine-tune DialoGPT models. The fine-tuning will iterate over the training data. The pretrained model will use the context as input and make predictions. Then the loss between predictions and actual responses of restaurants will be calculated. Based on the loss, the parameters of the pretrained model will be updated. This process will be repeated and the checkpoints which record the model configuration are saved. The validation set will be used to select the checkpoint which performs the best on unseen data. Finally, we can obtain a fine-tune model. The main libraries used in this task include Transformers and Torch. All the tasks in this thesis are completed in Python 3.10.12 environment provided by Google Colab.

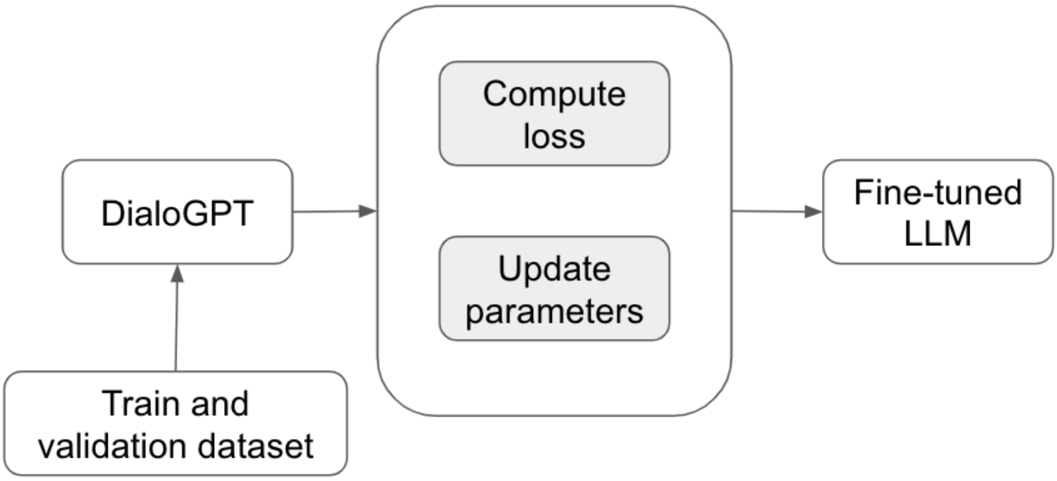


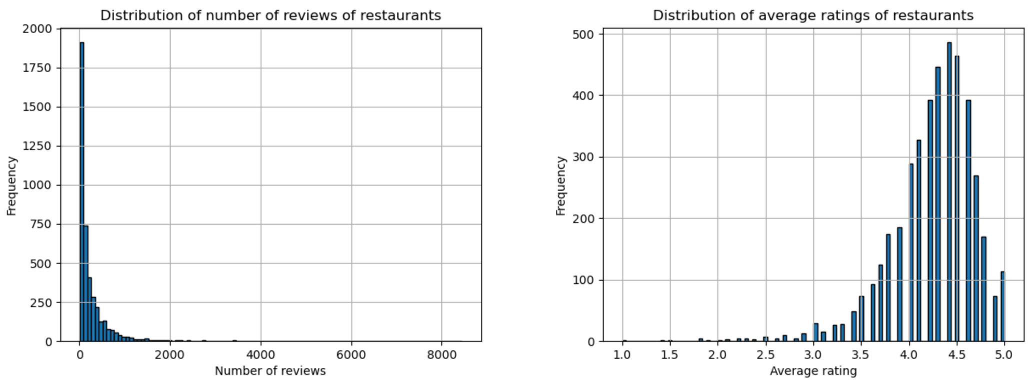
Figure 4. Fine-tuning process with domain-specific data.

The test set will be used for evaluation. BERTScore will be used as the metric for evaluating the generation results (T. Zhang et al., 2019). It compares the similarity between references and generated text by computing the similarity score of each token. It uses contextual embedding and performs better than other metrics which count n-gram overlaps. The value of BERTScore is between 0 and 1, and a larger value means the generated text and the reference are more similar. We will also calculate the BLEU (Bilingual Evaluation Understudy) scores of the generated text by the fine-tuned models. Although BLEU score is usually used to evaluate the quality of machine translation, it can be used as a reference in this study. It will calculate the proportion of words in the generated text that also appear in the reference responses of restaurants. A setting of BLEU-2 is applied, where "2" indicates that the precision of sequences of two consecutive words is evaluated. The values of BLEU scores also range from 0 to 1 and higher values means better translation quality.

4 Results

4.1 Statistics of customer reviews

In this section statistics of the customer review dataset will be presented to provide a general understanding of restaurant reviews. In Figure 5, the distributions of number of reviews and average ratings across the restaurants in the dataset are shown. In the dataset of this thesis, most restaurants get less than 500 reviews from customers. The maximum value of this number is 8450 while the minimum value is 1. For the distribution of average ratings, most ratings fall within the range from 3.5 to 5.



(a) *(b)*

Figure 5. Distributions of (a) number of reviews, (b) average rating.

On Google Maps customers can rate the restaurants from 1 to 5 stars. With this rating value, the reviews in the dataset can be assigned to five groups and comparison of these groups can be performed. Figure 6 shows the distribution of reviews with different ratings and the distribution of text reviews. It is found that the rating of around 60% of the reviews is 5 and 23% of the customers rate 4, indicating that most customers are satisfied or quite satisfied with their experience. 4.7% of the customers rate 1 and 78% of them leave text reviews. It is also noted that 71% of the customers who rate 2 write text reviews. For reviews with ratings 3 to 5, the proportions are not such high. Therefore, customers with low level of satisfaction are more likely to give text feedback.

Figure 6. Number of reviews for different values of rating.

Then we will explore the length of review text. Table 4 lists the mean, media, maximum, and minimum of the length of review text for each value of ratings. Note that we use the translated text provided by Google if a review is written in a language other than English. It is found that a review with a lower rating has a longer text on average. It is the same for the median length of reviews. There is a different trend for the maximum length that reviews with a 5-star rating have the longest text. All the groups have a minimum length of 1. Examples of these short reviews include “Slow”, “Okey”, and “Delicious”. The analysis of reviews with different ratings has been conducted. In the following sections, we will focus on understanding the reviews and providing solutions of improving customer service workflow with machine learning and generative AI.

Table 4. Length of reviews with different ratings

Rating	1	2	3	4	5
Mean	53.6	47	31.6	24.4	22.5
Median	33	28	17	14	13
Max	826	776	792	766	1183
Min	1	1	1	1	1

4.2 Topics in customer reviews

In this section, the results of exploration of customer reviews will be presented. In Google Maps, users can review a restaurant by selecting a rating and some of them would leave a text review. Out of the 1,135,436 reviews, 61% of them have text written by customers. LDA is applied to find the topics in the reviews. Using the iteration method introduced in Section 3.2, the optimal number of topics for this dataset is 3. Figure 7 is the distance map of the detected topics. In this figure, each bubble is a topic, and its size depends on its occurrence in all the reviews. It is found that there is no overlap between bubbles and the bubbles are evenly scattered across the chart, which means the model performs well. Of these bubbles, Topic 2 is the largest and includes 45.4% of the tokens in the dataset of review text, which means this topic has a high prevalence in all review text. The bar chart on the right shows the top 30 words in the reviews. It is noticed that "food" is mentioned the most frequently. The aspects of service and atmosphere are also in this top 30 list.

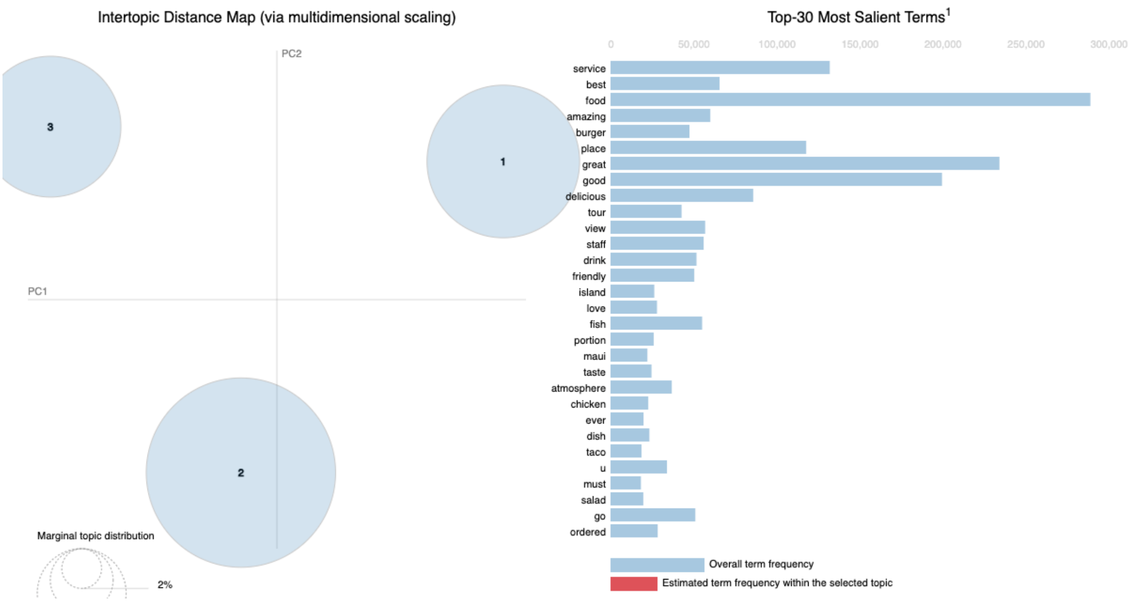


Figure 7. Distance among topics generated by the package pyLDAvis.

The word clouds of the top ten words in each topic are shown in Figure 8. The first topic is mainly about the experience of food. It mentions food related words, such as "delicious" and "taste", which refer to the aspect of food. Topic 2 mentions the words "service", "staff", and "friendly", which are all related to the aspect of service. Food and drink are also in the top ten list of this topic. Moreover, the word "view" is in the field of atmosphere, although it is not so frequent compared to service. Topic 3 mainly talks about some specific food, such as burgers and fish. In Figures A.2, A.3, and A.4 of appendices, the top 30 words of each topic are provided. It is noticed that the word “price”, which belongs to the aspect of value, is in the keywords of topic 1. From appendices, topic 3 is a mixed topic of food and atmosphere. Therefore, all important aspects that customers care about, as discussed in Section 3.2, have appeared in the topics detected by LDA. However, these aspects are not distributed throughout the topics alone, because customers may mention more than one aspect in one single review. It is also found that in all these identified topics, words related to food are of high frequency, indicating that customers of restaurants may care about the aspect of food the most. Moreover, several words of customers’ positive experience are in the keywords, e.g., “best”, “good” and “great”. This result is consistent with the one in the previous section that most customers are satisfied with the restaurants they visited.

Figure 8. Word clouds of the top ten words in each topic.

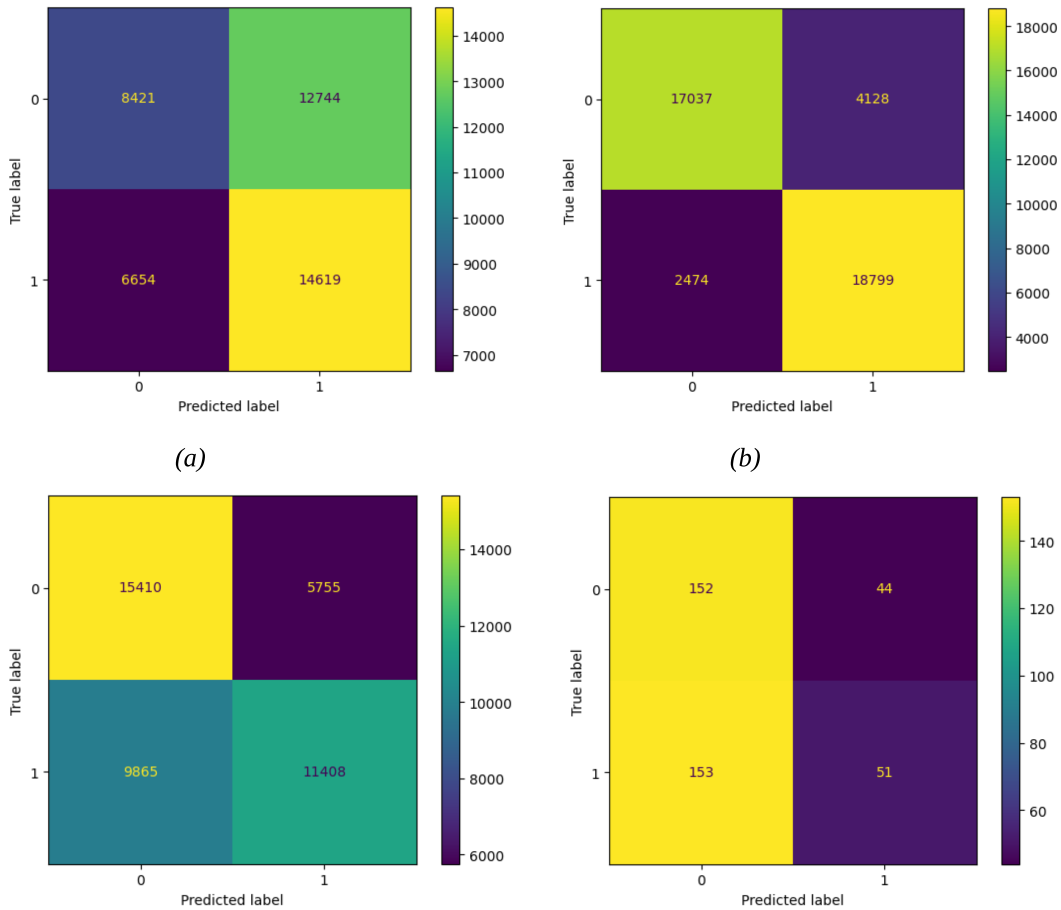
4.3 Classification of customer reviews

Four classification approaches are used in this section to model the prioritizing process of managers in deciding which reviews should get managerial responses. Table 5 lists the accuracy value for these approaches. In this table, the highest value of accuracy is from the random forest using numerical features. The neural network combining both numerical features and review text provides the accuracy value of 0.63, which is the second highest but poor performance. The other two approaches, logistic regression and sequential classification with BERT, have low accuracy.

Table 5. Accuracy of four approaches

Approach	Accuracy
Logistic regression	0.54
Random forest	0.84
Neural network	0.63
BERT	0.51

Figure 9 provides the confusion matrix for each approach. Logistic regression's predictions are positive for most cases, while BERT is likely to give negative predictions. Using the neural network approach, most predictions match the actual labels, but the quantity of erroneous predictions is still significant. In contrast, random forest's performance is much better. The incorrect predictions are much less than the correct ones. This result is consistent with the observation from Table 5.



(a) *(b)* *(c)* *(d)*

Figure 9. Confusion matrices for four approaches: (a) Logistic regression, (b) Random forest, (c) Neural network, (d) Sequential classification with BERT.

In Figure 10, the importance of features used in random forest is compared. It is noticed that the number of reviews of a restaurant is the most important feature in deciding whether to reply to a customer review. This figure also shows that the average rating of a restaurant would influence its intention to respond to its customer reviews. These two features are about restaurants. On the other hand, the features of a review itself, i.e. rating and whether there is text review, are not so important in this classification. Moreover, whether a review came with images does not have significant influence on the managerial response decisions. In summary, restaurants may exhibit considerable differences in the pattern of providing managerial responses. The features of reviews are of low importance. Therefore, results may vary if the classification approaches are applied to reviews of different restaurants.

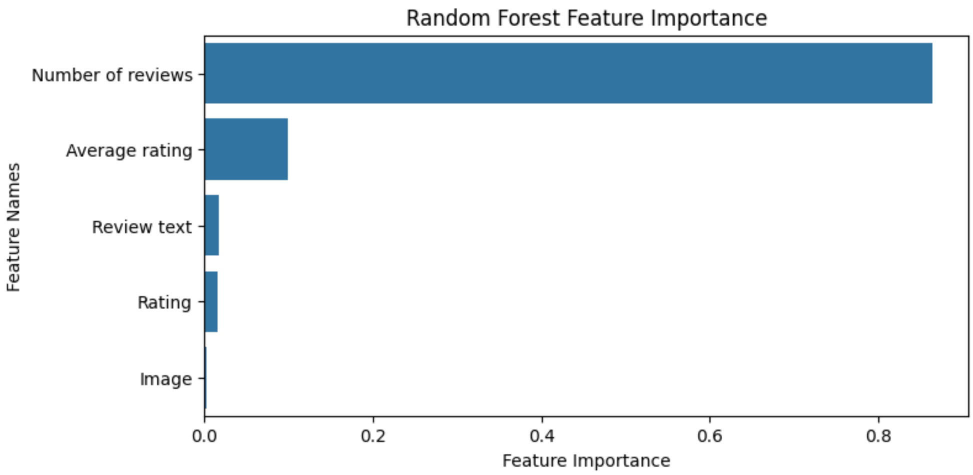


Figure 10. Feature importance of random forest.

To address this, we selected two restaurants and created two datasets of their reviews. These two restaurants are randomly selected from the restaurants with high volumes of reviews and responses. For the restaurants with only a few reviews, the data is not enough for model training. The above classification steps are repeated to verify whether the outcome is dependent on the dataset to which the model is applied. The results are listed in Table 6, and they are different from Table 5. For restaurant 1, all the four approaches show satisfying results of accuracy. Sequential classification using BERT performs the best. Although the neural network approach has a slightly lower accuracy, its speed is much faster than BERT. The results of restaurant 2 are quite different in that all classification methods show poor accuracy. The reason may be that managers of restaurants have their own criteria to write responses to customers. Some like to thank customers for rating high while some wish to explain to disappointed customers. It is also possible that some restaurants just reply to customers occasionally, without any criteria. The common implication of these two experiments is that including the review content makes classification better. Considering that using BERT is time consuming, it is suggested to try testing neural network first. Based on the accuracy, one can then decide whether to use it to classify the reviews which need to be replied to.

Table 6. Accuracy of four approaches in datasets of two restaurants

	Restaurant 1	Restaurant 2
Number of reviews	4821	3708
Average rating	4.3	4.3
Number of responses	1957	2090
Logistic regression	0.80	0.46
Random forest	0.80	0.54
Neural network	0.88	0.61
BERT	0.91	0.55

4.4 Generating managerial responses with AI

In this section, the results of fine-tuning large language models will be presented. In the context, names of reviewers, names of restaurants, ratings, and review text are included. With names of reviewers, the models can generate responses which add salutation by addressing the reviewers if needed. The names of restaurants can help identify different response styles and the ratings can provide some information about review sentiment. It is also possible that large language models can get sentimental information simply from review texts. However, not all the reviewers provide review text. Moreover, the ratings represent the total perceived values of customers, and the review text may not cover all the aspects. As restaurant managers would also refer to the rating values of reviews, we include this information in the context for fine-tuning. Large language models can make decisions on whether to use this information in generating.

BERTScore is used as the evaluation metric, and the scores for the two fine-tuned models are calculated in test set. Results are listed in Table 7. Both models show high scores, and DialoGPT-medium performs slightly better. Considering that DialoGPT-small is of small size and fine-tuning it takes less time, it is a good choice if one wants a model with good performance with limited resources. In Table 7 the BLEU scores are also provided. BLEU scores are not as high as BERTScores for the generated responses. It is reasonable because BLUE score is sensitive to word order and cannot capture semantic similarity between sentences. BLEU score for DialoGPT-medium is higher than the one for DialoGPT-small, which is consistent with the result of BERTScore.

Table 7. Evaluation of fine-tuned generative AI

	DialoGPT-small	DialoGPT-medium
BERTScore	0.9019	0.9077
BLEU score	0.4659	0.5194

The fine-tuned models can be applied to the test set and generate responses. The following Table 8 lists some examples of the test results by fine-tuned DialoGPT-small. For the first two examples, the managerial responses address the names of customers. The generated responses by the fine-tuned model also mention customers' names. This means that the large language model can understand the usage of the customer names provided in context. In the last example (e), the fine-tuned model utilizes the rating value in context to generate a response saying, "thank you for your Four Star Google Review". This is consistent with the reference response which also mentions this rating value.

The generated responses by the model can also imitate the writing styles of original responses by restaurants. In the first example (a), the response from manager contains some emoji characters, as so does the generated one. In example (b), the owner adds her name at the end of the response as a signature. The fine-tuned model has learned this pattern and does the same thing in generating a response. The restaurant of example (d) just replies in a short and quick way. The fine-tuned model also provides a short prediction. The responses of examples (b), (d), and (e) use the word "mahalo" in Hawaiian to express their appreciation. The fine-tuned model can generate responses to follow this usage.

However, this fine-tune model is still not perfect. For example, it cannot generate responses with specific details. In example (c) of Table 8, the customer is not satisfied with the food of the restaurant and complains about the sauce. The response from the restaurant expresses apology and explains the reason why the sauce is not satisfying. The model generated response apologizes and promises to forward the feedback to kitchen staff but does not offer explanations. Although the prediction of model is not a poor response, it cannot cover the details without related knowledge. Therefore, the participation of human is still needed in customer service. The good point is that the model does not make up a fake reason in this example.

Table 8. Examples of generated responses by fine-tuned model

	Rating	Review	Response	Prediction
(a)	5	Hawaii...with grandson visiting daughter stationed in Hawaii.	Sheila SCORE! 👏 Glad you spent your day the right way! 🍔🍹🕹️🏈	Sheila thanks for stopping by and for the review! Wouldn't have been as FUN without you! 🍔🍹🕹️🏀
(b)	4	None	Thanks for stopping by, Antonette! Come see us again soon. -Mahalo, Gina K., Owner	Thanks for coming by, Antonette! We're thrilled you're a fan. Come back to see us again soon. Mahalo. -Gina K., Owner
(c)	3	I ordered Teriyaki stir fry w/ chicken & noodles. The sauce was watery and I wasn't a fan of the noodles.	Sorry Jacob. It means the kitchen didn't drain your veggies properly. Sorry. The noodles are buckwheat soba Japanese noodles.	We are sorry you didn't enjoy your meal. We will share your feedback with our kitchen and management staff. Thank you for your time to share your experience.
(d)	5	Eat here at least twice every visit to Kona. Great atmosphere, food and music	Mahalo!	Mahalo for coming in 🤙
(e)	4	(Translated by Google) Spacious, clean and friendly (Original) 넓고 깨끗했고 친절했어요	Mahalo for the 4 star rating. We are glad that you found it comfortable!--CW	Aloha and thank you for your Four Star Google Review. Mahalo and we hope to see you again soon!

To compare the performance of general language models and fine-tuned models, we ask ChatGPT to help generate managerial responses. The same information is provided and ChatGPT is asked to act as a restaurant manager. In Table A1 of appendices, we list the responses from this model for the above reviews in Table 8. It is noticed that ChatGPT provides much more formal responses. And the lengths of the generated responses are almost the same. The generated responses of this model designed for general usage are also proper. However, it cannot imitate the linguistic styles of different restaurants. These responses are more likely from the same template. If a restaurant would like to keep its original replying style, a fine-tune model is better suited to meet the requirement.

5 Discussions

5.1 Discussion of results

In section 4, the experiment results have been presented. Statistics of the customer review dataset have shown that most reviews are from satisfied customers. Although the unhappy customers are more likely to write text reviews, they cover a small part of the dataset. Therefore, positive words appear frequently in the topics identified by LDA. The results also indicate that the aspects of a restaurant that customers emphasize include food, service, atmosphere, and price. This is consistent with previous research. It is noticed that food is the most important for these customers, and then followed by service and atmosphere. The words about price or value are not as frequent as the words related to other aspects. Such an observation is based on the dataset in this thesis. Results may be different if other datasets are used.

This thesis has attempted to classify customer reviews to decide which ones should be replied to. Four approaches are proposed to simulate the restaurants' decision process, and several features of reviews are used. Although the approach of random forest has a high value of accuracy, the important features are about restaurants but not about reviews. This means that differences exist between restaurants. The experiments are repeated in datasets of two randomly selected restaurants. However, the performance of all classification methods is poor in one of these restaurants. The reason may be the distinction of decision criteria between restaurants. Therefore, it does not work for all businesses to use machine learning methods to provide suggestions on whether a review needs a managerial response. It is suggested that making use of the text of reviews would provide better results.

The experiment of response generation with fine-tuned large language models has demonstrated the efficiency and quality of generative AI. After training with the reviews and responses in the dataset, large language models can generate responses which can identify and make use of specific information provided in context. Moreover, the models can imitate the writing styles of different restaurants. Nevertheless, the models cannot answer some questions not included in training data. If a restaurant does not have enough samples for models to learn, the generated responses may not be good enough.

5.2 Application of generative AI in customer service

One of the objectives of this thesis is to improve customer service procedure with generative AI. The results have verified the efficiency of managerial response generation for customer reviews. To apply generative AI in customer service, it is suggested to adjust application ways based on the specific conditions of each restaurant. For the restaurants with sufficient samples for model learning, e.g. thousands of review-response pairs, fine-tuning a model for each individual restaurant with its own dataset can generate responses better matching the style of its responses in the dataset. Then the model can be used to help reply to customer reviews. Before using the generated responses, managers or staff of the restaurant should examine them. If the generated responses contain inappropriate information, they should be revised. Restaurants can supplement details, such as explanations for questions and complaints.

For restaurants without enough samples, they can use a fine-tuned model for general restaurants to generate responses. The models in this thesis are models for general restaurants which use dataset of multiple restaurants in training. The performance of these models is good enough to improve the efficiency. Another approach is to train a model with dataset of multiple similar restaurants. The metrics to measure the similarity can be various. For example, restaurants within the same subcategory or restaurants with similar target customers can be comparable. The exploration of this approach can be done in future work. Also, participation of humans is needed for this kind of restaurants.

The application of classification is not recommended for all restaurants because the results of this thesis indicate that the classification processes are different among restaurants. The response decision criteria may be not constant but evolve over time even for an individual restaurant. Before training a classification model, it is required that there are enough data of this restaurant. If the data is imbalanced, more data are needed. For the restaurants which have enough data and a set of standards to evaluate the reviews and make decisions, making use of machine learning to improve this process may be possible. Moreover, the classification models should be updated if the standards are altered. The application of this requires deeper analysis.

5.3 Ethical considerations in the application of generative AI

There are some ethical considerations for the use of generative AI in customer service. First, the bias of training data should be considered. Large language models are trained on large datasets which usually contain biases. In this thesis, large language models are fine-tuned using the responses and reviews in our dataset. If the training data of the original models and our domain-specific data are biased, the unfairness may be represented in the generated contents. A possible solution is to ensure that the fine-tuning data is diverse and representative. The participation of humans, as it is suggested in this thesis, can also weaken the possible biases.

The second issue is transparency in the use of generative AI. Informing customers that the responses are generated with the assistance of AI helps strengthen the trust with customers. However, customers may have concerns if they have negative perception of AI technologies. Moreover, customers may be afraid that their opinions are not received by the management teams. In the previous section, managers are suggested to examine the generated responses before using them. We suggest emphasizing the participation of humans while informing the use of generative AI. By claiming that the reviews are read, and response contents are reviewed by humans, customers can be reassured.

Privacy is also an ethical issue in the use of generative AI. The data from customers should be handled in compliance with relevant laws and regulations. It should be ensured that customers consent to how their data are used. Moreover, it is suggested to formulate guidelines to avoid the misuse of generative AI.

In summary, the application of generative AI in customer service is feasible under the supervision of humans. Further research in this area would provide better and more detailed approaches of application. Moreover, the application of generative AI should consider ethical issues and comply with local laws and regulations. This thesis only uses dataset of restaurants, but other industries can also gain insight from the results of this thesis.

6 Conclusion and future works

This thesis has studied the customer reviews of restaurants and discussed the possibility of applying generative AI in improving customer service efficiency. The basic statistics of the dataset are presented and LDA is applied to provide a primary understanding of customer reviews. This study has tried to figure out if machine learning helps reduce the efforts in deciding reviews which need responses. Although the results suggest that review classification is difficult to apply, this study also provides some suggestions. The application of generative AI to generate responses by fine-tuning large language models is demonstrated to be feasible. The fine-tune models can generate responses with high BERTScore values, which calculate the similarity between generated and reference responses. Moreover, the generated responses can imitate the writing styles of the original ones well. Suggestions on how different restaurants apply generative AI are provided in the discussion section.

This thesis is a preliminary study of the application of generative AI. Future works can include the investigation of tailored models for restaurants without sufficient training data. Timestamps of reviews and responses can also be considered in future works to explore whether topics in review and response patterns evolve over time. The current study is based on the customer reviews of restaurants. Other categories of businesses can be included in future research.

References

Abdullah, R., Suhariyanto, & Sarno, R. (2021). Aspect based sentiment analysis for explicit and implicit aspects in restaurant review using grammatical rules, hybrid approach, and SentiCircle. *International Journal of Intelligent Engineering and Systems*, *14*(5), 294–305. https://doi.org/10.22266/IJIES2021.1031.27

Aguinis, H., Beltran, J. R., & Cope, A. (2024). How to use generative AI as a human resource management assistant. *Organizational Dynamics*, *53*(1), 101029. https://doi.org/10.1016/J.ORGDYN.2024.101029

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Castillo, D., Canhoto, A. I., & Said, E. (2021). The dark side of AI-powered service interactions: exploring the process of co-destruction from the customer perspective. *The Service Industries Journal*, *41*(13–14), 900–925. https://doi.org/10.1080/02642069.2020.1787993

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Deng, C., & Ravichandran, T. (2023). Managerial Response to Online Positive Reviews: Helpful or Harmful? *Information Systems Research*. https://doi.org/10.1287/ISRE.2019.0175

Dogru, T., Line, N., Mody, M., Hanks, L., Abbott, J., Acikgoz, F., Assaf, A., Bakir, S., Berbekova, A., Bilgihan, A., Dalton, A., Erkmen, E., Geronasso, M., Gomez, D., Graves, S., Iskender, A., Ivanov, S., Kizildag, M., Lee, M., … Zhang, T. (2023). Generative Artificial Intelligence in the Hospitality and Tourism Industry: Developing a Framework for Future Research. *Journal of Hospitality & Tourism Research*. https://doi.org/10.1177/10963480231188663

Flavián, C., Pérez-Rueda, A., Belanche, D., & Casaló, L. V. (2022). Intention to use analytical artificial intelligence (AI) in services – the effect of technology readiness and awareness. *Journal of Service Management*, *33*(2), 293–320. https://doi.org/10.1108/JOSM-10-2020-0378/FULL/PDF

Gan, J., & Qi, Y. (2021). Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. *Entropy 2021, Vol. 23, Page 1301*, *23*(10), 1301. https://doi.org/10.3390/E23101301

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467–483. https://doi.org/10.1016/J.TOURMAN.2016.09.009

Gutierrez-Bustamante, M., & Espinosa-Leal, L. (2022). Natural language processing methods for scoring sustainability reports—A study of Nordic listed companies. *Sustainability*, *14*(15), 9165.

Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International journal of environmental research and public health*, *18*(1), 271.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Jia, S. (Sixue). (2020). Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews. *Tourism Management*, *78*, 104071. https://doi.org/10.1016/J.TOURMAN.2019.104071

Jiménez, F. R., & Mendoza, N. A. (2013). Too popular to ignore: The influence of online reviews on purchase intentions of search and experience products. *Journal of interactive marketing*, *27*(3), 226-235. https://doi.org/10.1016/J.INTMAR.2013.04.004

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories: analysis of ChatGPT. *Central European Management Journal*, *31*(1), 3–13. https://doi.org/10.1108/CEMJ-02-2023-0091/FULL/PDF

Kudeshia, C., & Kumar, A. (2017). Social eWOM: does it affect the brand attitude and purchase intention of brands? *Management Research Review*, *40*(3), 310–330. https://doi.org/10.1108/MRR-07-2015-0161/FULL/PDF

Li, J., Shang, J., & McAuley, J. (2022). UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 6159–6169. https://doi.org/10.18653/V1/2022.ACL-LONG.426

Li, M., & Wang, R. (2023). Chatbots in e-commerce: The effect of chatbot language style on customers' continuance usage intention and attitude toward brand. *Journal of Retailing and Consumer Services*, *71*, 103209. https://doi.org/10.1016/J.JRETCONSER.2022.103209

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, *29*(3), 458–468. https://doi.org/10.1016/J.TOURMAN.2007.05.011

Liu, C. H. S., Su, C. S., Gan, B., & Chou, S. F. (2014). Effective restaurant rating scale development and a mystery shopper evaluation approach. *International Journal of Hospitality Management*, *43*, 53–64. https://doi.org/10.1016/J.IJHM.2014.08.002

Liu, C., Jiang, J., Xiong, C., Yang, Y., & Ye, J. (2020). Towards Building an Intelligent Chatbot for Customer Service: Learning to Respond at the Appropriate Time.

Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 3377–3385. https://doi.org/10.1145/3394486.3403390

Liu, S., Law, R., Rong, J., Li, G., & Hall, J. (2013). Analyzing changes in hotel customers' expectations by trip mode. *International Journal of Hospitality Management*, *34*, 359–371. https://doi.org/10.1016/J.IJHM.2012.11.011

Liu, S., Peng, C., Wang, C., Chen, X., & Song, S. (2023). icsBERTs: Optimizing Pre-trained Language Models in Intelligent Customer Service. *Procedia Computer Science*, *222*, 127–136. https://doi.org/10.1016/J.PROCS.2023.08.150

Liu, X., & Law, R. (2019). Insights into managers' response behavior: Priority and effort. *International Journal of Hospitality Management*, *77*, 468–470. https://doi.org/10.1016/J.IJHM.2018.08.010

Mathwick, C., & Mosteller, J. (2017). Online reviewer engagement: A typology based on reviewer motivations. *Journal of Service Research*, *20*(2), 204–218. https://doi.org/10.1177/1094670516682088

Olujimi, P. A., & Ade-Ibijola, A. (2023). NLP techniques for automating responses to customer queries: a systematic review. *Discover Artificial Intelligence 2023 3:1*, *3*(1), 1–19. https://doi.org/10.1007/S44163-023-00065-5

Panchendrarajan, R., Ahamed, N., Murugaiah, B., Sivakumar, P., Ranathunga, S., & Pemasiri, A. (2016). Implicit Aspect Detection in Restaurant Reviews using Cooccurence of Words. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 128–136. https://doi.org/10.18653/V1/W16-0421

Piskorski, J., Haneczok, J., & Jacquet, G. (2020). New Benchmark Corpus and Models for Fine-grained Event Classification: To BERT or not to BERT?. *In Proceedings of the 28th international conference on computational linguistics* (pp. 6663-6678).

Pournader, M., Ghaderi, H., Hassanzadegan, A., & Fahimnia, B. (2021). Artificial intelligence applications in supply chain management. *International Journal of Production Economics*, *241*, 108250. https://doi.org/10.1016/J.IJPE.2021.108250

Prentice, C., Dominique Lopes, S., & Wang, X. (2020). The impact of artificial intelligence and employee service quality on customer satisfaction and loyalty. *Journal of Hospitality Marketing & Management*, *29*(7), 739–756. https://doi.org/10.1080/19368623.2020.1722304

Qi, J., Zhang, Z., Jeon, S., & Zhou, Y. (2016). Mining customer requirements from online reviews: A product improvement perspective. *Information & Management*, *53*(8), 951–963. https://doi.org/10.1016/J.IM.2016.06.002

Qiu, K., & Zhang, L. (2023). How online reviews affect purchase intention: A meta-analysis across contextual and cultural factors. *Data and Information Management*, 100058. https://doi.org/10.1016/J.DIM.2023.100058

Ren, X., Wang, L., & Luo, X. (Robert). (2024). How customized managerial responses influence subsequent consumer ratings: The language style matching perspective. *Decision Support Systems*, *180*, 114188. https://doi.org/10.1016/J.DSS.2024.114188

Rhee, H. T., Yang, S. B., & Kim, K. (2016). Exploring the comparative salience of restaurant attributes: A conjoint analysis approach. *International Journal of Information Management*, *36*(6), 1360–1370. https://doi.org/10.1016/J.IJINFOMGT.2016.03.001

Rodler, S., Ganjavi, C., De Backer, P., Magoulianitis, V., Ramacciotti, L. S., De Castro Abreu, A. L., Gill, I. S., & Cacciamani, G. E. (2024). Generative artificial intelligence in surgery. *Surgery*. https://doi.org/10.1016/J.SURG.2024.02.019

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and Tourism Online Reviews: Recent Trends and Future Directions. *Journal of Travel & Tourism Marketing*, *32*(5), 608–621. https://doi.org/10.1080/10548408.2014.933154

Serra Cantallops, A., & Salvi, F. (2014). New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management*, *36*, 41–51. https://doi.org/10.1016/J.IJHM.2013.08.007

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, *32*(6), 1310–1323. https://doi.org/10.1016/J.TOURMAN.2010.12.011

Tran, L. T. T. (2020). Online reviews and purchase intention: A cosmopolitanism perspective. *Tourism Management Perspectives*, *35*, 100722. https://doi.org/10.1016/J.TMP.2020.100722

Van Noort, G., & Willemsen, L. M. (2012). Online Damage Control: The Effects of Proactive versus Reactive Webcare Interventions in Consumer-generated and Brand-generated Platforms. *Journal of interactive marketing*, *26*(3), 131–140. https://doi.org/10.1016/J.INTMAR.2011.07.001

Verma, R. K., & Kumari, N. (2023). Generative AI as a Tool for Enhancing Customer Relationship Management Automation and Personalization Techniques. *International Journal of Responsible Artificial Intelligence*, *13*(9), 1–8. https://neuralslate.com/index.php/Journal-of-Responsible-AI/article/view/66

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, *30*(1), 123–127. https://doi.org/10.1016/J.TOURMAN.2008.04.008

Wang, P., Fang, J., & Reinspach, J. (2021). CS-BERT: a pretrained model for customer service dialogues. *NLP for Conversational AI, NLP4ConvAI 2021 - Proceedings of the 3rd Workshop*, 130–142. https://doi.org/10.18653/V1/2021.NLP4CONVAI-1.13

Wu, J., Ye, J., Zhou, K., & Chen, L. (2023). To respond or not to respond? The reviewer- and review content-related influencers on managerial response decision towards customer reviews. *International Journal of Hospitality Management*, *114*, 103558. https://doi.org/10.1016/J.IJHM.2023.103558

Xie, K. L., Chen, C., & Wu, S. (2016). Online Consumer Review Factors Affecting Offline Hotel Popularity: Evidence from Tripadvisor. *Journal of Travel & Tourism Marketing*, *33*(2), 211–223. https://doi.org/10.1080/10548408.2015.1050538

Xie, K. L., Zhang, Z., Zhang, Z., Singh, A., & Lee, S. K. (2016). Effects of managerial response on consumer eWOM and hotel performance: Evidence from TripAdvisor. *International Journal of Contemporary Hospitality Management*, *28*(9), 2013–2034. https://doi.org/10.1108/IJCHM-06-2015-0290/FULL/PDF

Xu, X., & Zhao, Y. (2022). Examining the influence of linguistic characteristics of online managerial response on return customers' change in satisfaction with hotels. *International Journal of Hospitality Management*, *102*, 103146. https://doi.org/10.1016/J.IJHM.2022.103146

Yan, A., He, Z., Li, J., Zhang, T., & McAuley, J. (2023). Personalized Showcases: Generating Multi-Modal Explanations for Recommendations. *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, *5*(23), 2251–2255. https://doi.org/10.1145/3539618.3592036

Yu, X., Liu, Y., Huang, X., & An, A. (2010). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*(4), 720–734. https://doi.org/10.1109/TKDE.2010.269

Yu, D., & Xiang, B. (2023). Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. *Expert Systems with Applications*, 120114.

Zhang, M., Fan, B., Zhang, N., Wang, W., & Fan, W. (2021). Mining product innovation ideas from online reviews. *Information Processing & Management*, *58*(1), 102389. https://doi.org/10.1016/J.IPM.2020.102389

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *8th International Conference on Learning Representations, ICLR 2020*. https://arxiv.org/abs/1904.09675v3

Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2019). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30

Appendices

```
# Step 1: Pre-trained Model Selection
tokenizer = LOAD DialoGPT
model = LOAD DialoGPT model

# Step 2: Training Configuration
Arguments = SET Arguments (learning rate, batch size, epochs)
optimizer = SELECT Adam optimizer

# Step 3: Data Preparation
dataset = LOAD domain-specific review-response dataset
context = prepare context
Split the data into training_set, validation_set,test_set
tokenized_dataset = TOKENIZE and FORMAT data (input-output pairs)

# Step 4: Fine-tuning Process
INITIALIZE model with pre-trained weights
FOR epoch IN range(num_epochs):
  FOR batch IN training_set:
    outputs = model(FORWARD pass with inputs)
    loss = COMPUTE loss(outputs, actual_responses)
    gradients = BACKWARD pass(loss)
    optimizer.UPDATE weights(gradients)
    EVALUATE model on validation_set
  END FOR
END FOR

# Step 5: Evaluation
performance_metrics = EVALUATE on validation_set

# Step 6: Model Testing
test_performance = TEST model on separate test_set
```

Figure A.1. Pseudo-code for fine-tuning DialoGPT model.

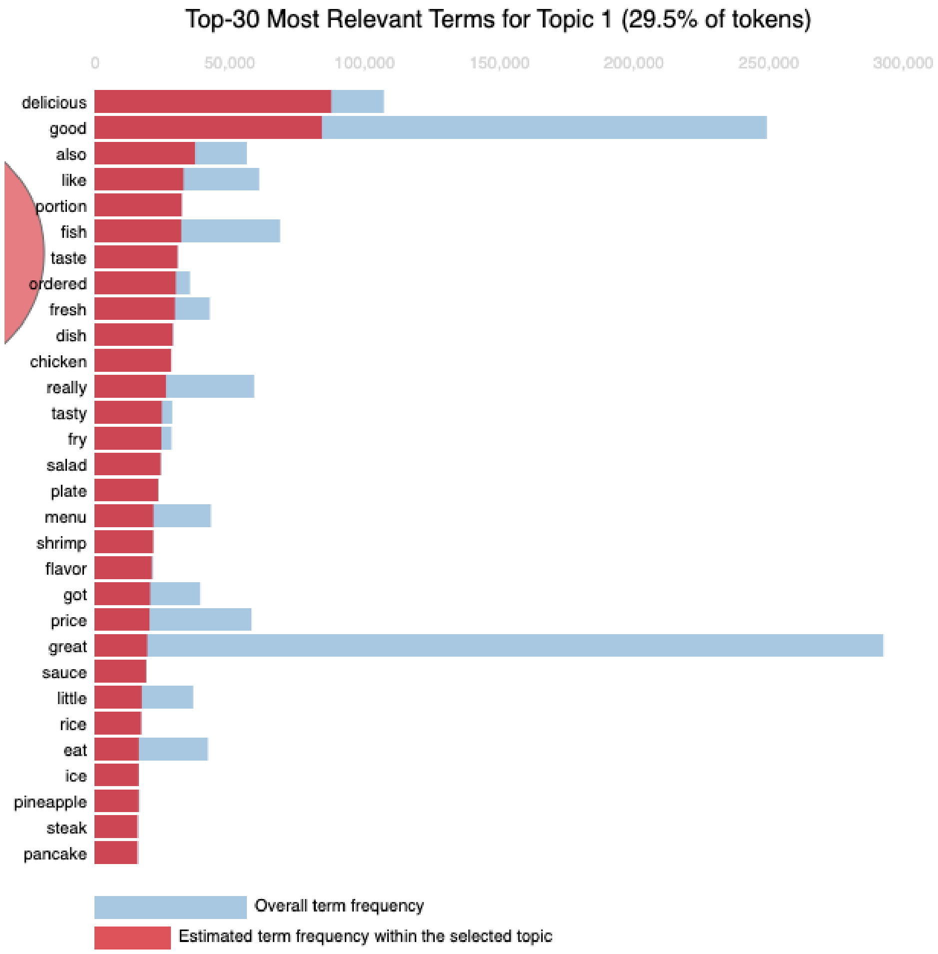


Figure A.2. Frequent words in topic 1 of the topic modelling result.

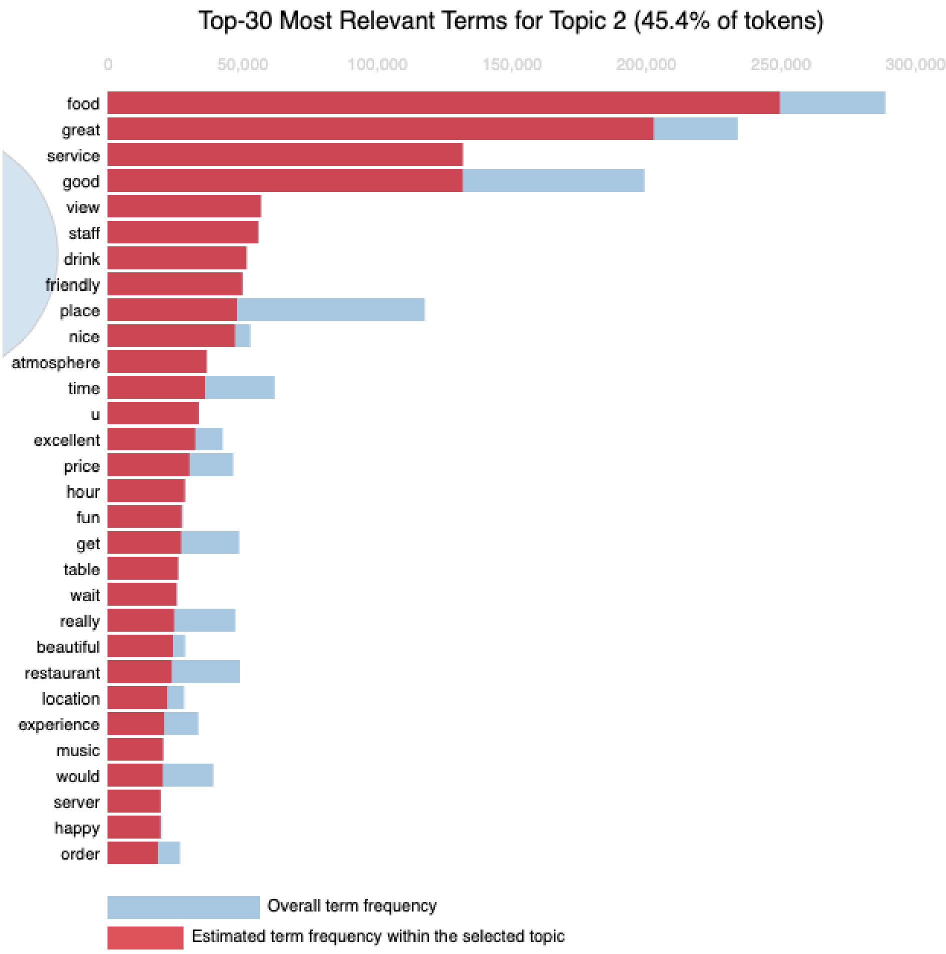


Figure A.3. Frequent words in topic 2 of the topic modelling result.

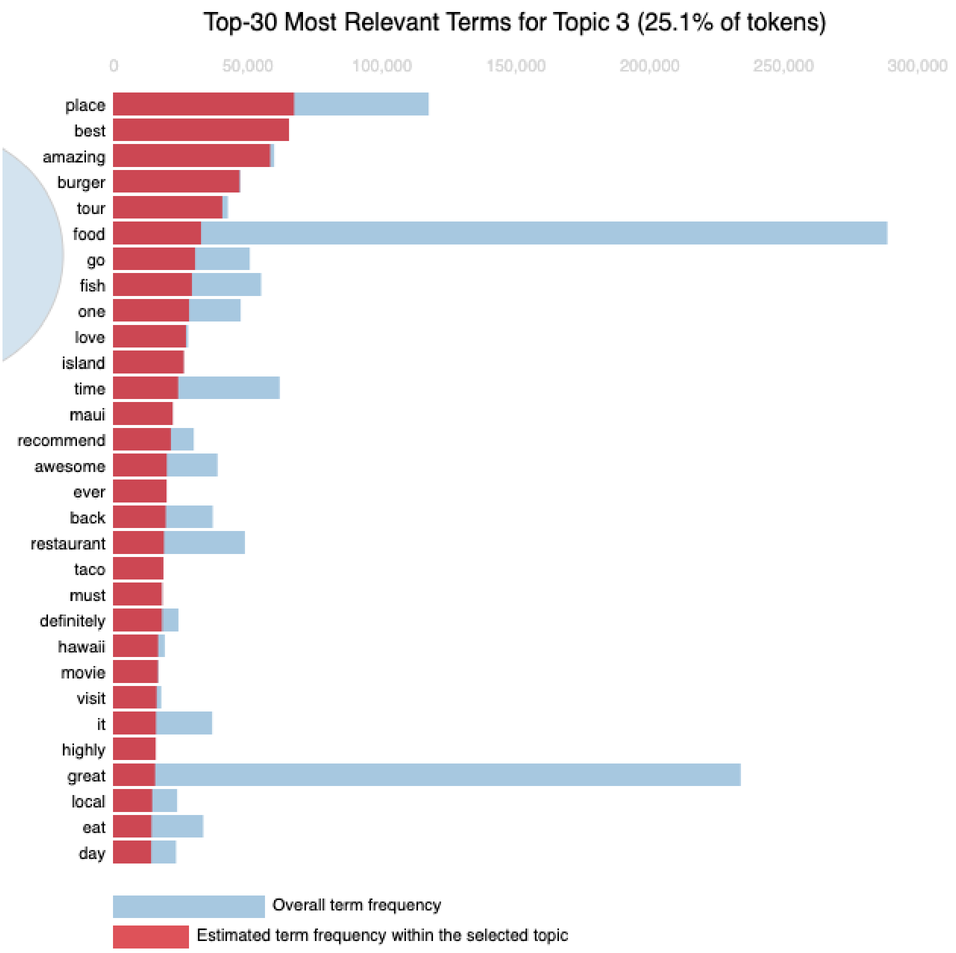


Figure A.4. Frequent words in topic 3 of the topic modelling result.

Table A.1. Examples of generated responses by ChatGPT

Rating	Review	Response	Prediction of ChatGPT
5	Hawaii...with grandson visiting daughter stationed in Hawaii.	Sheila SCORE! 👏 Glad you spent your day the right way! 🍔🍹🕹🏈	Hi Sheila, Thank you for your positive review and for choosing Dave & Buster's! We're thrilled to hear that you had a great time during your visit with your grandson. It's fantastic to know that our restaurant could contribute to your memorable experience. We appreciate your support and hope to see you again soon! Best regards, [Your Name]
4	None	Thanks for stopping by, Antonette! Come see us again soon. -Mahalo, Gina K., Owner	Hi Antonette, Thank you for taking the time to leave a review for Hawaiian Style Cafe - Waimea. We appreciate your feedback and are glad to hear about your experience. If you have any specific comments or suggestions for us, please feel free to share. We hope to have the opportunity to serve you again soon! Best regards, [Your Name]
3	I ordered Teriyaki stir fry w/ chicken & noodles. The sauce was watery and I wasn't a fan of the noodles.	Sorry Jacob. It means the kitchen didn't drain your veggies properly. Sorry. The noodles are buckwheat soba Japanese noodles.	Hello Jacob, Thank you for providing feedback on your recent visit to Bamboo Restaurant & Gallery. We apologize that your experience with our Teriyaki stir fry did not meet your expectations. We take your comments seriously and will address the issues with our culinary team to ensure consistent quality in our dishes. We hope you'll consider giving us another chance to impress you in the future. Best regards, [Your Name]
5	Eat here at least twice every visit to Kona. Great atmosphere, food and music	Mahalo!	Hi Lynn, Thank you for your glowing review and for being a loyal patron of On the Rocks! We're thrilled to hear that you enjoy our atmosphere, food, and music during your visits to Kona. Your support means the world to us, and we look forward to welcoming you back for many more memorable dining experiences. Warm regards, [Your Name]
4	(Translated by Google) Spacious, clean and friendly (Original) 넓고 깨끗했고 친절했어요	Mahalo for the 4 star rating. We are glad that you found it comfortable!--CW	Hello Jae Ho, Thank you for your review of Hard Rock Cafe! We're glad to hear that you found our establishment spacious, clean, and friendly. Customer satisfaction is our top priority, and we appreciate your positive feedback. We hope to see you again soon for another enjoyable dining experience. Best regards, [Your Name]