Personalizing Online Reviews for Better Customer Decision Making

by

© *Tao Chen*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

Faculty of *Business Administration*

Memorial University of Newfoundland

May 2020

St. John's Newfoundland

Abstract

Online consumer reviews have become an important source of information for understanding markets and customer preferences. When making purchase decisions, customers increasingly rely on user-generated online reviews; some even consider the information in online reviews more credible and trustworthy than information provided by vendors. Many studies have revealed that online reviews influence demand and sales. Others have shown the possibility of identifying customer interest in product attributes. However, little work has been done to address customer and review diversity in the process of examining reviews. This research intends to answer the research question: how can we solve the problem of customer and review diversity in the context of online reviews to recommend useful reviews based on customer preferences and improve product recommendation? Our approach to the question is through personalization. Similar to other personalization research, we use an attribute-based model to represent products and customer preferences. Unlike existing personalization research that uses a set of pre-defined product attributes, we explore the possibility of a data-driven approach for identifying more comprehensive product attributes from online reviews to model products and customer preferences. Specifically, we introduce a new topic model for product attribute identification and sentiment analysis. By differentiating word co-occurrences at the sentence level from at the document level, the model better identifies interpretable topics. The use of an inference network with shared structure enables the model to predict product attribute ratings accurately. Based on this topic model, we develop attribute-based representations of products, reviews and customer preferences and use them to construct the personalization of

online reviews. We examine personalization from the lens of consumer search theory and human information processing theory and test the hypotheses with an experiment. The personalization of online reviews can 1) recommend products matching customer's preferences; 2) improve custom's intention towards recommended products; 3) best distinguish recommended products from products that do not match customer's preferences; and 4) reduce decision effort.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In today's business environment, markets have become more saturated and competitive than ever, while customers are given unprecedented choices and information about products. Advances in information technology and the popularity of social networks have led to widespread customer sharing of product and service experiences. Such customer-generated online reviews provide valuable information about markets and customer preferences (Decker and Trusov, 2010) and become a de facto "sales assistant" to help customers to identify products that meet their needs (Chen and Xie, 2008). For businesses, due to the ease and the low cost with which large amounts of data can be collected, online review data are usually more comprehensive, representative and timely than data collected through traditional approaches such as expert panel, focus group and marketing survey. Many studies demonstrate the usefulness of online review data in revealing customer preferences towards products by showing that favourable overall ratings positively influence demand and sales across some product categories (e.g., Chevalier and Mayzlin, 2006; Cui et al., 2012;

Godes and Mayzlin, 2004; Liu, 2006; Onishi and Manchanda, 2012; Ye et al., 2009). For customers, online reviews are a new source of product information and have become an important means to reduce uncertainty about a product. A survey shows an increasing number of customers read online reviews before they make purchase decisions (Deloitte, 2014). A few studies investigate customers' perception of online reviews (Benlian et al., 2012; Mudambi and Schuff, 2010; Yin et al., 2014). Though these studies show online reviews can be valuable for both businesses and customers, they often simplify or abstract the rich text content of online reviews into, for example, overall rating/valence/sentiment, or overall helpfulness, disregarding various product attributes discussed in the text content and the possibility that certain product attributes are perceived differently from overall sentiment and helpfulness. *This research examines the question of how to solve the problem of customer and review diversity in the context of online reviews to recommend useful reviews based on customer preferences and improve product recommendation.* In answering the question, this research: 1) introduces a model to identify product attributes and the associated sentiments from the text content of online reviews, 2) develops a personalization system of online reviews to improve customer decision making, and 3) evaluates the effectiveness of the new personalization system.

A few studies recognize the importance of understanding customer preferences towards product attributes and focus on identifying product attributes from online reviews. Previous studies usually involve manual coding a large amount of text data (Godes and Mayzlin, 2004; Liu, 2006), and it would be impractical if product attribute identification is done similarly. Most recent studies have applied natural language processing (NLP) techniques to assist the process with more or less human

intervention. Two approaches have been used to identify product attributes. One approach is to use an NLP tagger to extract noun phrases or named entities and then, either manually, or using a lexical database such as WordNet (2016), or clustering algorithms, group the extracted noun phrases into product attributes (Archak et al., 2011; Decker and Trusov, 2010; Netzer et al., 2012). The other approach is to apply clustering algorithms such as k-means, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or their variants, to the text content; each cluster identified by the algorithm is considered as a product attribute (Lee and Bradlow, 2011; Tirunillai and Tellis, 2014). Neither approach is ideal. The first approach can easily miss the discussion of a product attribute if the noun phrases describing the product attribute are not explicitly mentioned, for example, "A little noisy in low light, for example on cloudy days, grass will lack sharpness and end up looking like a big mass of green." (Archak et al., 2011). This sentence discusses the image quality of a camera without using the noun phrases "image quality" and is most likely to be ignored. The second approach uses identified clusters to represent product attributes, but the identified clusters sometimes cannot be easily interpreted as product attributes, making the results unreliable. In addition, these studies either require the online reviews in a particular format to infer sentiments of product attributes, for example, the reviews must segregate the content into pros and cons sections, which significantly limits the application of the approaches (Decker and Trusov, 2010; Lee and Bradlow, 2011), or lack an effective method to identify the sentiments that understands multi-word negative structures and idiomatic expressions (Archak et al., 2011; Netzer et al., 2012; Tirunillai and Tellis, 2014).

This research introduces a novel topic model to address these problems. The

model identifies topics at both the sentence level and the document level and produces high-quality topics that are suitable for interpretation as product attributes. The model uses an inference network to establish an explicit link between the hidden product attribute sentiment variables and the review text, which enables the model to understand multi-word negative structures and idiomatic expressions and improves the performance of predicting product attribute ratings.

This research is closely related to studies of web personalization and recommendation (e.g., Chen et al., 2009; Ho and Bodoff, 2014; Liang et al., 2006; Tam and Ho, 2005). Tam and Ho (2005) use the Elaboration Likelihood Model to examine the relationship between web personalization and customer persuasion. Both Liang et al. (2006) and Chen et al. (2009) apply the theory of information overload to identify the relationships among web personalization, information overload and customer decision-making. Ho and Bodoff (2014) extend the Elaboration Likelihood Model with Consumer Search Theory to study the customer's product screening process under the web personalization to understand customer attitude. Similarly, this research is interested in understanding the effect of personalization on customer behaviour, but is different from traditional personalization and recommendation studies in two ways: the content of personalization and the approach to modelling customer preferences. First, existing studies have focused on personalizing the offering of products matching the customer's preferences. Although this research is also interested in understanding customer preferences, the focus is to personalize online reviews and examine how personalized online reviews affect customer decision-making. Second, in existing recommendation and personalization studies, customer preferences are modelled by a few pre-defined attributes of customers and products, usually identified

by the researchers. The studies in predicting customer rating have shown modelling preferences based on a set of pre-defined attributes does not perform well (Bell et al., 2010), suggesting the set of pre-defined attributes usually cannot account for customer diversity. This research uses a data-driven approach to identify product attributes from online reviews and can identify a more comprehensive set of product attributes and better model customer preferences.

This research conceives a view of products, online reviews and customers based on product attributes, utilities of product attributes and customer preferences over product attributes, as illustrated in Figure 1.1. A product consists of a set of product attributes. Each product attribute carries an intrinsic utility. The aggregate of the product attribute utilities determines the utility of the product. An online review evaluates a few product attributes of the product and conveys the intrinsic utility of the evaluated product attribute through the discussion of the product attribute. The discussion of product attributes forms the complete review and is abstracted as the overall rating. Since a customer has no direct access to the utilities of product attributes of a product, the customer needs to examine reviews to estimate the expected utility of product attributes and the expected utility of the product based on the customer's preferences. Often, the customer needs to dig through tons of less useful reviews to come to an inaccurate estimation and make a suboptimal decision. In this research, we propose a model to reveal the intrinsic utility of product attributes and customer preferences through online reviews. We design a personalization approach using the uncovered intrinsic utilities of product attributes and customer preferences to improve recommendations and customer decision-making.

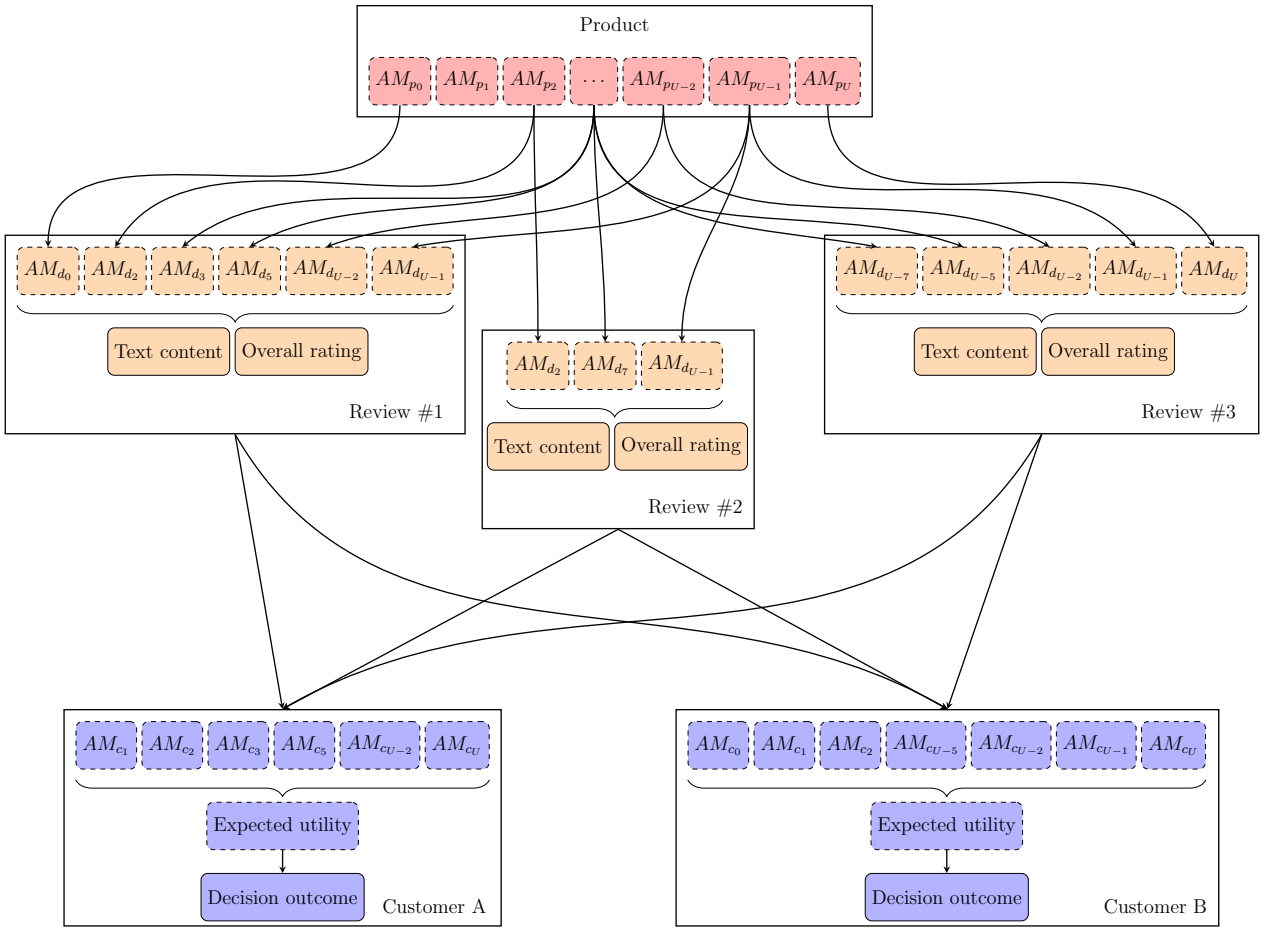

Figure 1.1: Products, Online Reviews and Customers: $AM_{p(\cdot)}$ is a product attribute of the product that carries an intrinsic utility; $AM_{d(\cdot)}$ is the conveyed utility of the product attribute through the review; $AM_{c(\cdot)}$ is the customer's preference on the product attribute. All of them are of the dashed boxes, indicating that they are unobservable. The observable variables are the text content and the overall rating of online reviews and the decision outcomes of customers. Our research proposes models to learn the hidden variables and uses them to improve customer decision making.

The contributions of this thesis are four-fold. First, it answers the research question by introducing a new way of personalizing online reviews. Second, it develops a new topic model that identifies coherent topics that are easy for interpretation and predicts product attribute rating accurately, to enable the design of the personalization of online reviews. Third, it analyzes the personalization of online reviews under different types of products and using different sorting mechanisms. Finally, it implements the personalization design and empirically confirms the hypotheses.

The rest of the thesis is organized as follows. Chapter 2 reviews the related literature and provides the motivation of this research in relation to the research of online review, personalization and recommendation, and sentiment analysis on product attributes. Chapter 3 introduces the topic model for product attribute identification and sentiment analysis. Chapter 4 discusses customer preference modelling, preference matching of products and online reviews, and the development of personalization features on online reviews. Chapter 5 presents the experiment design to evaluate the personalization of online reviews and discusses the empirical results. Chapter 6 highlights and summarizes the contributions of this research.

Chapter 2

Literature Review

This chapter outlines the journey that inspires this research. This research starts with the online reviews literature in Section 2.1 that helps us understand the effect of online reviews on the individual customer's decision-making. However, the diversity in customers and online reviews is often missed from the discussion. To have a better understanding of customer diversity on decision-making, we turn to the studies of personalization in Section 2.2, which provide the theories that support the idea of using personalization to address customer diversity, improve decision-making and reduce decision effort. Meanwhile, the theories offer guidance for designing the personalization of online reviews. The sentiment analysis technique on product attributes is the enabler of the personalization of online reviews. Section 2.3 reviews the existing techniques, identifies the problems in them that impact the effectiveness of the personalization of online reviews, and discusses our approach to address these problems.

2.1 Online Reviews

Online reviews and online word-of-mouth have received considerable attention in marketing and information system research. One of the foci is to understand the effect of online reviews on aggregated customer behaviour. Many studies examine the correlation between online reviews and product sales. For example, Godes and Mayzlin (2004), and Liu (2006) find online reviews have explanatory power over TV show ratings and box office revenue. Many studies confirm that online review sentiments or ratings are positively correlated to product sales across different product categories (Chevalier and Mayzlin, 2006; Chintagunta et al., 2010; Clemons et al., 2006; Cui et al., 2012; Dellarocas et al., 2007; Ye et al., 2009; Zhu and Zhang, 2010). A few investigate the effect of online reviews on product sales under the interaction with, for example, traditional media (advertising) (Onishi and Manchanda, 2012) and competing products (Jabr and Zheng, 2014; Kwark et al., 2014), and also find positive relationships between review sentiments and product sales. Besides review sentiments, researchers identify the effect of other aspects of online reviews, such as reviewer characteristics (Forman et al., 2008), trust (Pavlou and Dimoka, 2006), and online review system design (Jiang and Guo, 2015), on product sales or pricing. Closely related to this research, Chen et al. (2008) find review quality, measured by helpfulness votes, has a positive effect on product sales. A series of studies closely examined review helpfulness and found review extremity, review length (Mudambi and Schuff, 2010), linguistic style (Cao et al., 2011; Schindler and Bickart, 2012), statement types (descriptive and evaluative) (Schindler and Bickart, 2012), and emotion (Yin et al., 2014) contribute to perceived review helpfulness. Recent studies (Hu and

Chen, 2016; Siering et al., 2018; Singh et al., 2017; Zhang and Lin, 2018) look beyond explaining into predicting review helpfulness, in which they confirm the effect of the previously identified factors, and discover the influence of reviewer characteristics, such as reviewer's reputation and expertise, on review helpfulness.

Though contributing to deepening the understanding of online reviews, these studies do not take into account a common phenomenon that one review may discuss multiple product attributes and convey different sentiments towards different product attributes. Given this observation, a few studies manage to extract product attributes from online reviews automatically by employing various NLP techniques. Decker and Trusov (2010), though do not report the detailed technique used for product attribute extraction, discuss the challenge in identifying product attributes in reviews, review the existing approaches and outline their procedure in collecting product attributes. They extract words and phrases that are interpretable as product attributes, determine their sentiments based on the pros or cons sections that the words and phrases appear, and merge the synonyms as one product attribute. By mapping the review text to product attributes and sentiments, they conduct regression analysis of the effect of product attributes on the overall evaluation of the products to understand the aggregated customer preferences. Lee and Bradlow (2011) use the k-means clustering algorithm to identify product attributes from the review text. They compare the automatically identified product attributes from reviews to the expert-defined product attributes, and use a survey to confirm the automatically identified product attributes are interpretable and understandable. They demonstrate the automatically identified product attributes from reviews can be used to analyze the market structure and produce a similar result as the traditional market research method does. One limitation

of both studies is that they require the reviews to have separated pros and cons sections to determine the sentiments of the product attributes. This requirement makes such approaches inapplicable to the many reviews that do not have such a format.

Archak et al. (2011) use the Part-of-Speech (POS) tagger to identify frequently mentioned noun phrases and apply WordNet and a hierarchical clustering algorithm to organize identified noun phrases to product attributes. They use dependency parser to extract the adjectives that modify the noun phrases of product attributes as the customer opinions towards product attributes, and propose two approaches to either cluster the opinion phrases into a few categories or map the opinion phrases into sentiment polarity scores. They compare the automatically identified product attributes to the crowdsourcing identified product attributes and the product attributes identified by two human annotators, and confirm the interpretability and significance of the automatically identified product attributes. They demonstrate the significant explanatory power of the product attributes and the customer opinions identified from review texts. Netzer et al. (2012) use the conditional random field (Lafferty et al., 2001) to develop a Named Entity Recognition model to identify product brands and models and a POS tagger to identify noun phrases representing customer complaints. They create a rule-based algorithm to fine-tune the results and group them into classes. Two human annotators validate the results in a sample to ensure the validity of the results. They use the extracted information to analyze the market structure to understand brand associations and to identify frequent complaints of products. Tirunillai and Tellis (2014) introduce a topic model to identify product attributes and corresponding sentiments. They examine the validity of interpreting the topics as product attributes by comparing to the results identified by human annotators and

the expert-defined product attributes. They demonstrate the possibility of tracking the market change over time with the product attributes and sentiments. Recently, a few studies follow the trend: they introduce a new approach that identifies product attributes and sentiments at the sentence level (Büschken and Allenby, 2016), and use the word-embedding method to prune sentences and group them into clusters to discover customer needs (Timoshenko and Hauser, 2019).

The discussion above shows an NLP process/model, which can identify or group words/phrases that have an interpretation of product attributes and can determine sentiments accurately, is essential for the research that considers the diversity of product attributes discussed in online reviews and uses an automated approach to identify product attributes and sentiments (Humphreys and Wang, 2017). This thesis reviews the existing methods of identifying product attributes and sentiments in Section 2.3. This research improves on these methods and introduces a new topic model for conducting a fine-grained analysis of product attributes and sentiments in online reviews to understand customer preferences over product attributes.

We notice these studies focus on the effect of online reviews on aggregated customer behaviour in the market but less on the effect on the individual customer. Though the studies of review helpfulness examine the individual's response to reviews, they dismiss the diversity in reviews and customers. Branco et al. (2012, 2015) examine the customer's purchase behaviour under the effect of online reviews. They view the customer's use of online reviews for decision-making as an information search process. In (Branco et al., 2012), they introduce an analytic model for the customer information search, from which they derive the optimum stopping rule for the information search and the purchase likelihood. They further analyze the

optimum pricing of the product and its impact on customer information search. In (Branco et al., 2015), they propose a different analytic model for the customer information search. Following a similar analysis, they examine how the amount of the product information provision affects the purchase decision-making. Though these two studies discuss the heterogeneous importance of product attributes to customers, their assumption that the customer can research the more important attributes earlier does not agree with the reality. Davis and Agrawal (2018) conduct an empirical study and propose the value-driven identification to capture the customer's perceived shared preferences of product attributes with a reviewer of a product. The study shows that value-driven identification leads to higher information adoption but does not explicitly identify product attributes and customer preferences, which limits the application of the results. We summarize recent studies of online reviews on product attributes in Table 2.1. This research focuses on the individual customer behaviour in the context of online reviews, explicitly identifies product attributes and customer preferences, and uses an information search view to understand the individual customer's decision-making under the personalization of online reviews. Section 2.2 reviews the current research in personalization and recommendation agents.

	Extract Product Attribute	Restrictions on Online Reviews	Research Objectives
Decker and Trusov (2010)	Yes	The review content must have separated the pros and cons sections	Derive aggregated customer preferences based on identified product attributes and sentiments from online reviews
Lee and Bradlow (2011)	Yes	As in (Decker and Trusov, 2010)	Analyze market structure based on identified product attributes and sentiments from online reviews
Archak et al. (2011)	Yes	No	Examine the explanatory power of the obtained product attributes and sentiments from online reviews on product sales

Netzer et al. (2012)	Yes	No	Analyze market structure and study brand associations using the extracted product attribute and sentiments from online reviews
Tirunillai and Tellis (2014)	Yes	No	Market change tracking with identified product attributes
Büschken and Allenby (2016)	Yes	No	Improve predicting power on product sales from product attributes and sentiments identified at sentence-level of online reviews
Timoshenko and Hauser (2019)	Yes	No	Reveal customer needs from identified product attributes in online reviews
Branco et al. (2012)	No	The customer can research the more important attributes earlier	Analyze the product pricing strategy in relation to the search cost in the context of online reviews
Branco et al. (2015)	No	As in (Branco et al., 2012)	Study the product information provision strategy in the context of online reviews
Davis and Agrawal (2018)	No	No	Examine how the shared preferences of product attributes between customers and reviewers affect customer information adoption

Table 2.1: Recent studies of online review on product attributes

2.2 Personalization

Online shopping can easily overwhelm customers with large volumes of information. An online marketplace, such as Amazon.com or Ebay.com, offers a large number of products. An abundance of product information and related reviews in such an online marketplace often strains human limits on attention and memory. Web personalization and recommendation agents, as a response to this challenge, have become popular in practice and research. Many studies investigate various aspects of web personalization and recommendation agents with different views and theories. Web personalization and recommendation agents are often viewed as a technological arti-

fact and, naturally, the technology acceptance model (TAM) (Davis, 1989) is often used to understand adoption (Koufaris, 2002). TAM is developed from the theory of reasoned action (Fishbein, 1979) in the context of information systems. It predicts and explains the acceptance and usage of information technologies based on two core constructs, i.e., perceived usefulness (PU) and perceived ease-of-use (PEOU). Perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance" and perceived ease-of-use is defined as "the degree to which a person believes that using a particular system would be free from effort." Other than PU and PEOU, many studies identify trust as an important construct in explaining technology adoption (Gefen et al., 2003; Komiak and Benbasat, 2006, 2008; Wang and Benbasat, 2016). Recent studies take a closer look at the design of personalization and recommendation. Ghoshal et al. (2015) introduce a recommendation technique using the proposed association rule method. The experiment shows their recommendations are more accurate than other existing approaches. Benlian (2015) investigates how the personalized content and design cues affect preference fit and further impact the adoption of the personalization and purchase decision. The author finds the content cues increase the preference fit that leads to an extended stay on the website and customers are more likely to make the purchase decision. Bernstein et al. (2018) present a clustering technique that effectively identifies customers with similar preferences and dynamically adjusts customer segmentation. Their case study shows that their approach substantially improves the transaction numbers.

Besides the view of technology adoption, Häubl and Trifts (2000) categorize a recommendation agent as an assistant for decision making and examine its effect

on decision quality. They find the use of a recommendation agent reduces cognitive efforts and leads to better decision quality. Tam and Ho (2006) examine web personalization from the view of human information processing, and find that personalization reduces decision effort and is more likely to influence choice when the recommendation matches customer's preferences well. Liang et al. (2006) and Chen et al. (2009) use the theory of information overload to examine the decision process. Their findings show web personalization reduces information overload and leads to better customer satisfaction and decision outcomes. Xiao and Benbasat (2007) integrate the view of technology adoption and decision-making processes in previous studies and propose a holistic conceptual model to understand the effect of recommendation agents on adoption, decision processes and decision outcomes. They also find that the use of a recommendation agent reduces decision effort and improves decision quality.

Tam and Ho (2005) conceptualize web personalization as a process of persuasion and use the Elaboration Likelihood model (Petty et al., 1983) to understand the effect of personalization on customer choice. ELM models the process of persuasive communications and posits two major routes to persuasion: the central route and the peripheral route. Persuasion under the central route involves a high level of elaboration and the persuasive message is more likely to influence attitude and predictably affect the decision. The level of elaboration under the peripheral route is low, the attitude change is less related to the content of the message but more to the recipient's current mood, and the decision becomes less predictable. The elaboration is defined as the extent to which a person carefully thinks about an argument. Tam and Ho (2005) find better preference matching of personalization leads to more elaboration and a higher chance of affecting customer choices. In a subsequent study, Ho and

Bodoff (2014) examines not only the persuasion process but also the product screening process and their interaction. They extend the Elaboration Likelihood Model by distinguishing three aspects of attitude: valence, persistence, and confidence. They posit the attitude confidence is related to the number of recommendations the customer samples. The more recommendations the customer samples the more confident the customer becomes. As the customer becomes confident, the customer stops examining more recommendations where they use the consumer search theory to determine the stopping time. They find the customer has more confidence in the product and is more likely to make the purchase decision if the recommendations become more and more fit for the customer's needs.

Regardless of the different theoretical lenses, these studies demonstrate that personalization contributes to better decision-making and less decision effort when the recommendation matches the customer's preferences. This research is based on these studies to guide the design of the personalization of online reviews: the personalization is to provide relevant and high-quality information to the customer to increase perceived usefulness, elaboration, attitude and confidence, decrease information overload, and enable the customer to improve decision-making and reduce decision effort. The primary contribution of this research is that, while previous studies focus on the personalization of the product offering, this thesis is interested in the personalization of online reviews matching customer's preferences. The main observation is that the amount of information in online reviews is much more than in product descriptions. Consequently, it requires significantly more cognitive effort from the customer to search and evaluate the information. As the search cost is higher in examining online reviews than in examining product descriptions, the customer is more likely

to give up or be ill-informed, which ends up with wasted efforts or a poor decision for the customer, and loss of sales for the business. The personalization approach of online reviews introduced in this research intends to better inform the customer with the relevant information in which the customer is interested, thus reducing the search cost and decision effort, and improving the information quality, and ultimately leading to better decision-making. The other difference is that prior research often models that customer preferences with a set of pre-defined product attributes in these studies (e.g. Ho and Bodoff, 2014; Tam and Ho, 2005, 2006), which may not account for the customer diversity and results in poor preference matching and unreliable results (Bell et al., 2010). In contrast, this research uses a data-driven approach to identify a more comprehensive set of product attributes to better modelling customer preferences and reliably predicting customer ratings.

2.3 Sentiment Analysis on Product Attributes

Sentiment analysis is used to determine the opinion expressed in a piece of text. Turney (2002) and Pang et al. (2002) started the early work of applying machine learning techniques to automatic sentiment analysis. Since then, the area has burgeoned due to its wide commercial and political applications (Liu, 2015). Many studies focus on predicting the overall sentiment of a document and often treat this goal as a classification problem in machine learning (Pang and Lee, 2008). A large amount of available online reviews along with ratings provides the data required to train models. Many models use the bag-of-words representation and perform reasonably well when predicting the overall sentiment of a document. However, accuracy degrades

when analyzing at a more detailed level, for example, examining the sentiment of a sentence in a document.

Recent research starts focusing on fine-grained sentiment analysis. Socher et al. (2013) collect fine-grained sentiment data on sentences and sentence constituents and compose word-embedding and sentiment information into a parse tree. They employ the word-embedding representation that has become popular since then. Word-embedding (Bengio et al., 2003) is a different representation in NLP from the bag-of-words representation. The bag-of-words representation models a piece of text as a vector and each component of the vector corresponds to a word in the vocabulary. The value of the component is usually the number of occurrences of the word in the text. Word-embedding representation models a word with a fixed length vector and the value of the vector is learned from data. The representation of a longer text is constructed from the vectors corresponding to the words in the text and a function that reduces the sequence of vectors to a vector of the same length. If a word conveys the same meaning as a longer text, the distance between the vectors representing the word and the text is expected to be small. One constraint in the study is the requirement of collecting detailed sentiment data, which are scarce and expensive to acquire. Several studies apply the concept of multi-instance learning (Dietterich et al., 1997) to ease the constraint on data, viewing a document as a bag of sentences where each sentence may convey different sentiments from the overall sentiment of the document. They show performance improvement in predicting the sentiments of sentences (Kotzias et al., 2015; Pappas and Popescu-Belis, 2014). However, multi-instance learning cannot address the problem of sentiment analysis on product attributes as the approach does not inform the product attributes discussed in a sentence.

In the literature, the problem of sentiment analysis on product attributes is usually termed as multi-aspect sentiment analysis (MASA) (Liu, 2012). It is worth mentioning that aspect-based sentiment analysis (ABSA) is a close but different type of sentiment analysis from MASA. ABSA works with a set of pre-defined aspects. ABSA models require the data consists of not only the overall sentiments but also the aspect sentiments (Pontiki et al., 2016), but such data is much less available. In MASA, the aspects can be pre-determined or not, and the data only has the overall sentiments but does not consist of the aspect sentiments. As the constraint on the data is relaxed, the MASA models have wider applications than the ABSA models. As the ABSA models have access to more information than the MASA models, the state-of-the-art ABSA model defines a performance upper bound for the MASA models for the same dataset. In this research, we compare our MASA model not only to the other MASA models but also to the ABSA models to have a better understanding of the performance.

Multi-aspect sentiment analysis inherently involves two tasks: opinion aspect extraction and sentiment analysis on the extracted opinion aspects. One approach to extract opinion aspects is to develop syntactical rules. A well-known syntactical approach is called "double propagation" (Qiu et al., 2011). The key idea is opinion words often modify aspect words and conjunction words often join two aspect words together. Some typical rules are "a noun phrase modified by an opinion word (such as good and great) is considered to be an aspect word (such as screen size and battery life for cell phone)," "an adjective modifying an aspect word is considered to be an opinion word," and "the noun phrase is considered to be an aspect word if a conjunction word joins the noun phrase with another aspect word." The rules

are applied to the review data repeatedly populating both aspect words and opinion words. The extracted aspect words and opinion words require further selection by evaluating against labelled data (Liu et al., 2015), or checking similarity in word-embedding and inspecting support and confidence using association rules (Liu et al., 2016). One advantage of this approach is that it only needs a small set of seed words to work appropriately, compared to the large labelled data set that most classification algorithms require. However, it still needs to group the extracted aspect words with similar meaning into product attributes, determining the sentiment polarity of the newly extracted opinion words, and conducting further sentiment analysis at least at the sentence level.

Besides the syntactical approach, many studies use the statistical topic model approach, such as the LDA model (Blei et al., 2003) and its variants (e.g., Wang et al., 2011; Zhao et al., 2010) to extract opinion aspects. These models correspond to the identified topics in the topic model to the aspects and simultaneously predict the associated sentiments of these topics. For example, Wang et al. (2011) use a regression approach that is similar to the supervised LDA (Blei and Mcauliffe, 2008) to model the overall rating. The regression model in the supervised LDA uses the sum of the topic assignment of all words as the regressor. Wang et al. assume each word has a sentiment weight for each topic. The word topic assignments and the word sentiment weights for the topic define the aspect sentiments, and the aspect sentiments define the overall rating. As a result, their regression model has two levels that the top level is a regression of the overall rating from the aspect sentiments and the bottom level is a regression of the aspect sentiments from the word topic assignments and the word sentiment weights for the topic. The approach assumes the opinion words that

describe the sentiment of a topic are also belong to the same topic. The assumption may not hold well in the topic model as the topic model assigned words to the same topic based on co-occurrences. The same opinion word may modify noun phrases of different topics. The co-occurrences of the opinion words and the noun phrases do not determine the topic assignment of the opinion words. In the other hand, the opinion words that carry similar sentiments often co-occur, for example, "good" and "great" or "bad" and "terrible". The opinion words are more likely to form their own topics instead of being grouped into the topic that they modify.

To address the topic assignment problem of the opinion words, Jo and Oh (2011) enforce one topic for a sentence instead of one topic for a word in LDA, and model the sentiment as a latent variable that influences the word generation from the topics. They use a few positive and negative seed words, and place weights in the prior of the word sentiment weight for topics accordingly, such that the latent sentiment variable can correspond to the positive and negative sentiments. Their approach of assigning one topic for a sentence can address the misalignment of the opinion words and the noun phrases in the topics as the noun phrases and the opinion words that modify them are often in the same sentence and are forced to be one topic. But the approach leads to poor performance in topic identification as the topic of a sentence is more likely to be dominated by the uninformative common words, for example, "best" or "ever" in "best taco ever" rather than "taco", due to their high frequency in reviews. It becomes unreliable to correspond to the topics determined by such common words to a ratable aspect. The similar problem also exists in other studies (e.g., Büschken and Allenby, 2016). Also, their approach cannot map the latent sentiment variable to more than two sentiments, for example, predicting a 5-star rating for an aspect.

Titov and McDonald (2008a) notice the problem that the topics identified by the LDA model and others do not correspond to ratable aspects. They introduce the global and local topics where each word in a sentence is either from a global topic of the document or a local topic of the sentence or the surrounding sentences. They show that some of the local topics can be interpreted as ratable aspects, for example, the "rooms", "service" and "location" aspects of hotels in hotel reviews. The use of global topics and local topics improves the identification of the ratable aspects and increases the reliability of the aspect sentiment prediction. Other studies use the similar the technique to extract ratable aspects (e.g., Wang et al., 2016; Zhao et al., 2010). Our model also distinguishes the global and local topics to model the topics at the sentence level better and correspond the topics to ratable aspects reliably. But, our model, different from Titov and McDonald's approach, examines the scope of word co-occurrences, either inside a sentence or between sentences, and explores the relationship between global and local topics.

A common conceptualization of a review document in multi-aspect or aspect-based sentiment analysis is to view a review as a bag of opinion phrases where each opinion phrase is a pair of aspect noun phrases and opinion words. A group of studies subscribe to the view and try to extract opinion phrases from reviews. Zhao et al. (2010) extend the LDA topic model to extract opinion phrases. They posit a sentence in a review consists of three types of words: aspect words, opinion words and background words, where different types of words form different types of topics and the aspect words and opinion words together form the opinion phrases. To distinguish the aspect words and the opinion words, they pre-train a maximum entropy model (Ratnaparkhi, 1996) that relies on the syntactic information, such as Part-of-Speech,

to classify two types of words. The probability output from the maximum entropy model feeds into the LDA model to assign words into different types. The LDA model further assigns the words to different topics. They demonstrate the approach is capable of extracting aspect words and associated opinion words. But, in terms of predicting aspect sentiments, the approach falls short as the topic model is not suitable to understand the multi-word negations.

To achieve better performance in predicting aspect sentiments, Moghaddam and Ester (2011) use a dependency parser to convert a sentence to an opinion phrase where the negations become a unique token, for example, the phrase "not bad" becomes a token "not_bad" that is unrelated to the word "bad". They use two LDA models to represent the aspect words and the opinion words in an opinion phrase. One LDA model organizes different aspect words that describe the same aspect to one aspect topic, and the other LDA model groups different opinion words that convey the same sentiment to one sentiment topic. There are two limitations of the approach. First, as the dependency parser converts the negations to a token unrelated to the original opinion word, the model at least needs to have both the original opinion word and the negative opinion word in the training data to understand that the negative opinion word has the opposite sentiment of the original word. For example, if the training data only consists of the word "good", the model would not understand "not_good" even if there exist other opinion words of the same negations. Secondly, they set the number of sentiment topics for the opinion words the same as the rating scales, for example, five sentiment topics for the 5-star rating scale. This design requires interpretation of the sentiment topics from the word assignment to determine the corresponding ratings, which can be hard to distinguish among close ratings. Wang and Ester

(2014) attempt to address the sentiment topic and rating correspondence problem by using an external sentiment lexicon consisting of opinion words and polarity scores. The assumption is that polarity scores correspond to ratings, and the experiment shows positive results in aligning sentiment topics to ratings. However, their model is unable to address the first problem.

Recent studies recognize the limitation of the topic model in understanding the multi-word negations and idiomatic expressions conveying sentiments. As the review data usually consist of reviewer and product information, these studies use the collaborative filtering approach for better sentiment rating prediction (Cheng et al., 2018; Wu and Ester, 2015). For example, Wu and Ester (2015) view the overall rating as the dot product of the aspect weights and the aspect sentiments in the review. They model the aspect weights and the aspect sentiments as a function of products and reviewers. The function for the aspect weights reflects whether an aspect is often mentioned when reviewing the product and whether the reviewer is inclined to discuss an aspect when reviewing products. The function of the aspect sentiments reflects the reviewer's preference over an aspect. The introduction of reviewers and products to the model helps improve the aspect rating prediction performance for the existing products and reviewers but does not benefit new products or new reviewers. Cheng et al. (2018) further show the improved performance in predicting overall ratings from introducing an extra layer of complexity that maps the aspect sentiments to the unexplainable latent factors. These approaches are less useful for the personalization as the improved performance does not come from a better understanding of the review text and the personalization requires the review text to provide evidence for decision-making.

We summarize the recent studies of sentiment analysis on product attributes in Table 2.2. Our model uses global and local topics and models the relation between them to improve the topic interpretability. Our model uses a similar formulation of the overall rating in terms of the aspect weights and the aspect sentiments but allows the model to use more topics than the number of pre-defined product attributes. It aims to improve the reliability of interpreting topics as ratable product attributes and to better account for the overall ratings from the discussion beyond pre-defined product attributes. We innovatively use the inference network to constrain the aspect sentiments to the corresponding text. The use of the inference network allows the model to leverage the recent development of word embedding and deep convolution network (Gehring et al., 2017) that is capable of understanding the multi-word negations and idiomatic expressions.

	Require product attribute ratings	Parsing sentences to aspect-opinion phrase pairs	Interpretability of identified product attributes	Comprehensiveness of identified product attributes	Use of overall rating	Predict product attribute rating	Product attribute rating alignment with the rating scale	Support multi-word idiomatic expressions in sentiment analysis	Support negations in sentiment analysis
Titov and McDonald (2008a)	Yes	No	Good. The model distinguishes global and local topics and treats local topics as ratable product attributes. Empirical evidence shows that local topics map to ratable product attributes well	No	No	Yes	Yes. The model requires product attribute ratings in training data to establish the correspondence.	Yes	No. Max-Entropy model, being a linear model, has trouble to recognize unseen negations
Zhao et al. (2010)	No	No	Good. The model uses a pre-trained Max-Entropy model to determine whether a word is an aspect word. The topics only consisting of aspect words are easy to interpret	No	No	No	N/A	N/A	N/A

Wang et al. (2011)	No	No	Poor. The model confines the number of topics to the number of pre-defined ratable product attributes. Using a rather small number of topics forces the model to merge the discussion of less related product attributes together, which may reduce the reliability when interpreting a topic as a ratable product attribute	No	Yes	Yes	No. The model maps the rating distribution to a normal distribution, but the unconstrained sentiment weights may distort the alignment	No	No
Jo and Oh (2011)	No	No	Poor. The model forces all words in a sentence to have the same topic. The common but less meaningful words tend to decide the topic, which leads to unreliable interpretation	No	No	Yes	No	No	No
Moghaddam and Ester (2011)	No	Yes	Depends on how well the parsing algorithm identifies the aspect phrases in sentences	No	No	Yes	No	Depends on how well the parsing algorithm identifies the opinion phrases in sentences	No

Wang and Ester (2014)	No	Yes	Depends on how well the parsing algorithm identifies the aspect phrases in sentences	No	No	The model predicts ratings for products instead of reviews	Depends on how well the sentiment lexicon corresponds to the rating scale	Depends on how well the parsing algorithm identifies the opinion phrases in sentences	No
Bagheri et al. (2014)	N/A	No	Good. The model uses a HMM model to capture multi-word expressions in topics. The multi-word expressions help the interpretation of the topics as ratable product attributes	Yes	N/A	N/A	N/A	N/A	N/A
Wu and Ester (2015)	No	No	Poor. The model has the same interpretation issues as in (Wang et al., 2011)	No	Yes	Yes	No. The model has the same unconstrained sentiment weight issue as in (Wang et al., 2011)	No	No
Cheng et al. (2018)	No	No	Poor. The model has the same interpretation issues as in (Wang et al., 2011)	No	Yes	No	N/A	No	No

Table 2.2: Recent studies of sentiment analysis on product attributes

2.4 Summary

In this chapter, we review the literature on online reviews and identify the gap in the knowledge of online reviews and personalization. We are interested in designing a personalization of online reviews to assist customer decision-making. The review of the literature in personalization provides the theories to guide the design of the personalization and use a data-driven approach to identify a much comprehensive set of product attributes for better preference matching. The review of the recent development of sentiment analysis on product attributes identifies the key problems that our model should address: better modelling the topics in a sentence and conduct sentiment analysis at the sentence level that is capable of understanding negations and idiomatic expressions.

Chapter 3

Attribute-Sentiment Analysis Sentence Model

In this chapter, we introduce a novel topic model that differentiates two types of topics in the context of online reviews to better identify product attributes. We integrate the sentiment analysis component to the topic model for multi-aspect sentiment analysis. We evaluate the performance of the model in identifying topics and in predicting product attribute ratings. The results show that the model has advantages over existing models and is suitable to support the development of the personalization of online reviews.

3.1 Modelling Sentences in Online Review

Topic models such as the LDA model have many successes in modelling text collections such as news articles, scientific publications, and Wikipedia web pages (Hoffman et al., 2013). However, such topic models are less effective when applied to user-

generated online reviews for multi-aspect sentiment analysis for two reasons. First, user-generated reviews are much shorter than the articles in the text collections mentioned above, or expert reviews. User-generated reviews often use short sentences discussing different product attributes. The chance of co-occurrence of related words in a sentence is much lower. For example, in a financial news article, it is common to see co-occurrence words such as "bank," "dollar" and "stock" in one sentence. However, a review of a Mexican restaurant may mention "best fish tacos ever" in a sentence and then moves to the discussion of the location in the next sentence. The review lacks the co-occurrence of "taco" and other Mexican food in a sentence. Secondly, multi-aspect sentiment analysis requires understanding the product attributes discussed in a sentence rather than the overall product attribute distribution of the whole review as the product attributes in the discussion change with the sentences and the sentiments change as well. A review may compliment the food of a restaurant in one sentence but criticize the service in the next sentence. The overall product attribute distribution cannot distinguish the sentiment difference between food and service.

One approach used in multi-aspect sentiment analysis is to model the overall product attribute distribution of the review but assign one topic for all words in a sentence (Jo and Oh, 2011). An issue of this approach is that common words are likely to dominate the topic of the sentence and may fail to identify the product attribute of the sentence. For example, a sentence "best fish tacos ever" is more likely to be assigned to a topic identified by the words "best" and "ever" rather than "fish tacos" since "best" and "ever" are much common than "fish tacos". Our approach distinguishes two types of topics and combines them to define the topic distribution

of a sentence of a review. We call one type of topics the inter-sentence topic and the other type the intra-sentence topic. For the intra-sentence topic, the related words are likely to co-occur in the same sentence or share some common words with which they co-occur in sentences. For the inter-sentence topic, the related words are less likely to co-occur in the same sentence but more likely in the same review.. For example, the words "taco" and "burrito" in the restaurant reviews are more likely to form an inter-sentence topic as they may not co-occur in the same sentence but are often seen together in the same review. The words "wait" and "minutes" are more likely to form an intra-sentence topic as they often appear in the same sentence. In this model, each sentence of a review in the model has its intra-sentence topic variables. The intra-sentence topic variable organizes the related words into one topic by connecting the word re-occurring in different sentences with the words often co-occurring with the re-occurring word in different sentences. At the same time, all sentences of the review share one inter-sentence topic variable. The inter-sentence topic variable links the words that are less likely to co-occur in the same sentence but often co-occur in the same review to one inter-sentence topic.

Besides the distinction of two types of topics, the shared inter-sentence topic variables are unlikely to carry the same weight for every sentence; some sentences may focus on "Mexican food" while others may focus on "wait time". One approach is to define scaling variables applied to the inter-sentence topic variables element-wise to reflect the various degrees of prevalence of the inter-sentence topic in a sentence. Our approach comes from the observation that some intra-sentence topics co-occur with inter-sentence topics while others do not. For example, a sentence of the intra-sentence topic "food taste" often contains words from the inter-sentence topic "Mex-

ican food" but a sentence of the topic "wait time" is less likely to consist of words associated with the topic "Mexican food". Our model maps the intra-sentence topics variables to the scaling variables for each sentence and applies the scaling variables to the inter-sentence topics to determine the contribution of the inter-sentence topics in the sentence. The idea of distinguishing two types of topics is related to the global and local topics in (Titov and McDonald, 2008b,a), but we model the topics and words based on the scope of co-occurrences and explore the interaction among two types of topic variables. Our model also better models the variation of the inter-sentence topics in the sentences and reduces the number of topic variables to avoid over-parameterization, which leads to the degradation of performance in topic identification (Tang et al., 2014).

Notation	Description
$\odot$	Point-wise product operator
$\oplus$	Vector concatenation operator
D	Dataset of reviews
d	Review document
t	Sentence in a review document
$\|t\|$	Number of words in a sentence t
V	Number of words in the vocabulary
w	Word in the vocabulary
T	Number of sentences in a review document
k^s	Intra-sentence topic
k^p	Inter-sentence topic
k, k'	Intra or inter-sentence topic
K^s	Number of intra-sentence topics
K^p	Number of inter-sentence topics
K	Total number of topics where $K = K^s + K^p$
θ_d^p	Inter-sentence topic distribution of a review d
θ_t^s	Intra-sentence topic distribution of a sentence t
θ_t	Full sentence topic distribution consisting of inter and intra sentence topics of size K defined as $\theta_t = \left(\mathrm{Q}(\theta_t^s) \odot \theta_d^p\right) \oplus \theta_t^s$

Notation	Description
W	Topic word distribution of shape $(V \times K)$ and its row entry W_w is the topic distribution of a word w
Q	Mapping function of shape $(K^p \times K^s)$ that maps the intra-sentence topic distribution to scaling variables determining the effect of the inter-sentence topic distribution in a sentence
$n(w,t)$	Word count of a word w in a sentence t
α_w, β_w	Hyper-parameter of the prior Gamma distribution of the topic word distribution
α_q, β_q	Hyper-parameter of the prior Gamma distribution of the mapping function Q
α_c^p, β_c^p	Hyper-parameter of the prior Gamma distribution of the inter-sentence topic distribution
α_c^s, β_c^s	Hyper-parameter of the prior Gamma distribution of the intra-sentence topic distribution
u	Ratable product attribute in the sentiment analysis
U	Number of ratable product attributes in the sentiment analysis
$\mathrm{I}(\cdot)$	Indicator function
r	Overall rating of a review d
R	Overall ratings of the dataset D
$\mathrm{k}(u)$	Mapping function that maps a ratable product attribute u in the sentiment analysis to a topic k
r^b	Residual sentiment rating
r_u^a	Sentiment rating of a ratable product attribute u in the sentiment analysis
$\lambda_r(r^b, r^a_{0\ldots U}, \theta_{0\ldots T})$	Poisson regression function for the overall rating of a review
J	Size of the word-embedding

Table 3.1: Mathematical notation

The model relies on the sparsity property of the Gamma distribution for modelling topic distributions and topic word distributions (Ranganath et al., 2015) and uses Poisson distributions to generate words. Based on the above discussion, we formulate the topic distribution of a sentence as the concatenation of two components: the adjusted inter-sentence topics and the intra-sentence topics, and define the topic

distribution of a sentence as follows,

$$\theta_t = \left(\mathrm{Q}(\theta_t^s) \odot \theta_d^p \right) \oplus \theta_t^s \tag{3.1}$$

We summarize the generative story of the model as follows. Table 3.1 lists the notations used in the discussion.

- Generate a topic word matrix W such that each entry $W_{wk} \sim \text{Gamma}(\alpha_w, \beta_w)$
- Generate a mapping function Q such that each entry $\mathrm{Q}_{k_1 k_2} \sim \text{Gamma}(\alpha_q, \beta_q)$
- For each review $d \in D$,
    1. Generate the inter-sentence topic θ_d^p where each entry $\theta_{dk^p}^p \sim \text{Gamma}(\alpha_c^p, \beta_c^p)$
    2. For each sentence t in the review d,
        - Generate the intra-sentence topic θ_t^s where each entry $\theta_{tk^s}^s \sim \text{Gamma}(\alpha_c^s, \beta_c^s)$
        - For each word w in the sentence t,
            * Compute the Poisson rate λ_w where $\lambda_w = W_w^\intercal \theta_t$
            * Generate the word count in the sentence $n(w,t) \sim \text{Poisson}(\lambda_w)$

Following the generative story, we have the formulation of the model as follows,

$$\begin{aligned} P(D|\alpha,\beta) \;=\; & \int d\,W\, d\,\mathrm{Q} P(W|\alpha_w,\beta_w) P(\mathrm{Q}|\alpha_c,\beta_c) \\ & \times \prod_{d\in D} \int d\,\theta_d^p\, P(\theta_d^p|\alpha_c^p,\beta_c^p) \\ & \times \prod_{t\in d} \int d\,\theta_t^s\, P(\theta_t^s|\alpha_c^s,\beta_c^s) \prod_{w\in t} P\Big(n(w,t) | W_w^\intercal \Big((\mathrm{Q}(\theta_t^s) \odot \theta_d^p) \oplus \theta_t^s \Big) \Big) \end{aligned}$$

We set α_c^s, α_c^p, α_w, and α_q to be less than 1 for sparsity. As a smaller value implies a sparser distribution, we set $\alpha_c^s < \alpha_c^p$ to reflect that a sentence is more likely to focus on

one local topic while a review may discuss many topics across sentences. As the exact inference of the posterior of the formulation is not tractable, we use the mean-field variational inference (Hoffman et al., 2013) to approximate the learning objective, from which we can derive a closed formed coordinated ascent algorithm. The detail of the derivation is in the Appendix A. Note that the attribute sentence model allows a closed formed optimization algorithm. The introduction of the sentiment analysis requires not only the mean-field variational inference but also the reparameterization technique (Kingma and Welling, 2014; Naesseth et al., 2017) to estimate the gradient for the optimization due to the normalization and the use of the indicator function in the attribute-sentiment analysis sentence model.

3.2 Integrating Sentiment Analysis

Commonly, the topics identified by the topic model may not support an interpretation as a ratable product attribute. Previous research usually forces the number of topics in the topic model to be equal to the number of designated aspects that are involved in the sentiment analysis (e.g. Wang and Ester, 2014; Wu and Ester, 2015), which constrains the model from fitting the data better and is likely to overflow the topics with irrelevant words leading to poor performance in predicting the aspect rating. Consider a sentence "Pool and hot tub were excellent" in a hotel review as an example. Assigning the sentence to any one of the six aspects, "Location," "Sleep Quality," "Room," "Service," "Value" and "Cleanliness" of the TripAdvisor dataset (TripAdvisor, 2015) would be questionable. Our attribute-sentiment analysis sentence model uses a much larger number of topics than the number of ratable

product attributes to accurately determine the existence of the discussion of product attributes in sentences. The model introduces a hidden sentiment rating r_u^a for each ratable product attribute u and a hidden residual sentiment rating r^b to represent the sentiments in the review that are not related to the ratable product attributes. The model conceives that a reviewer forms different sentiments toward different product attributes after assessing a product. The reviewer communicates the sentiments of some product attributes through the sentences discussing the product attributes, bags the sentiments of the remaining product attributes together and conveys through the sentences of general discussion. The sentiments of different product attributes reveal the overall rating, and the most discussed product attribute contributes the most to the overall rating. The sentiment rating for each product attribute represents the sentiments of the product attributes discussed in the sentences. The residual sentiment rating represents the bagged sentiment of the remaining product attributes conveyed through the general discussion. The generative story of the attribute-sentiment analysis sentence model is as follows, and its graphical representation is in Figure 3.1.

- Generate a topic word matrix W such that each entry $W_{wk} \sim \text{Gamma}(\alpha_w, \beta_w)$
- Generate a mapping function Q such that each entry $\text{Q}_{k_1 k_2} \sim \text{Gamma}(\alpha_q, \beta_q)$
- For each review $d \in D$,

  1-2. The same steps as in the attribute sentence model

  3. *Generate* r^b *where* $r^b \sim \text{Gamma}(r, 1)$

  4. *For each ratable product attribute* u,

     – *Generate* r_u^a *where* $r_u^a \sim \text{Gamma}(r, 1)$

5. *Generate an overall rating r where* $r \sim \text{Poisson}(\lambda^r(r^b, r^a_{1...U}, \theta_{1...T}))$

The italicized steps are the sentiment steps added to the attribute sentence model. The formulation of the attribute-sentiment analysis sentence model is,

$$\begin{aligned} P(D, R|\alpha, \beta) &= \int d\,W d\,\mathrm{Q} P(W|\alpha_w, \beta_w) P(\mathrm{Q}|\alpha_c, \beta_c) \\ &\times \prod_{d \in D} \int d\,\theta_d^p\, P(\theta^p|\alpha_c, \beta_c) \\ &\times \prod_{t \in d} \int d\,\theta_t^s\, P(\theta_t^s|\alpha_c, \beta_c) \prod_{w \in t} P\Big(n(w,t)|W_w^{\intercal}\theta_t\Big) \\ &\times \int d\,r^b P(r^b|r,1) \prod_u \int d\,r_u^a P(r_u^a|r,1) \\ &\times P(r|\lambda_r(r^b, r^a_{1...U}, \theta_{1...T})) \end{aligned}$$

The Poisson regression function $\lambda_r(r^b, r^a_{1...U}, \theta_{1...T})$ that generates the overall rating is defined as below,

$$\begin{aligned} \lambda_r(r^b, r^a_{0...U}, \theta_{0...T}) &= \frac{1}{\sum_t \sum_k \mathrm{I}(K\theta_{tk} > \sum \theta_{tk'})\theta_{tk}} \times \Bigg(\\ &\sum_t \sum_k \Big(\sum_u \mathrm{I}(\mathrm{k}(u) = k) r_u^a + \\ &(1 - \sum_u \mathrm{I}(\mathrm{k}(u) = k)) r^b \Big) \times \mathrm{I}(K\theta_{tk} > \sum \theta_{tk'})\, \theta_{tk} \Bigg) \end{aligned} \tag{3.2}$$

The regression function embodies the previous discussion of the relation between the overall rating, the sentence, and the sentiment rating of product attributes. A sentiment rating of a product attribute contributes to the overall rating in a sentence if a substantial number of words in the sentence are about the product attribute. If more words in the sentence are about the product attribute, more of the sentiment rating of the product attribute contributes to the overall rating in the sentence. If a sentence is less about product attributes but more about the general discussion, more

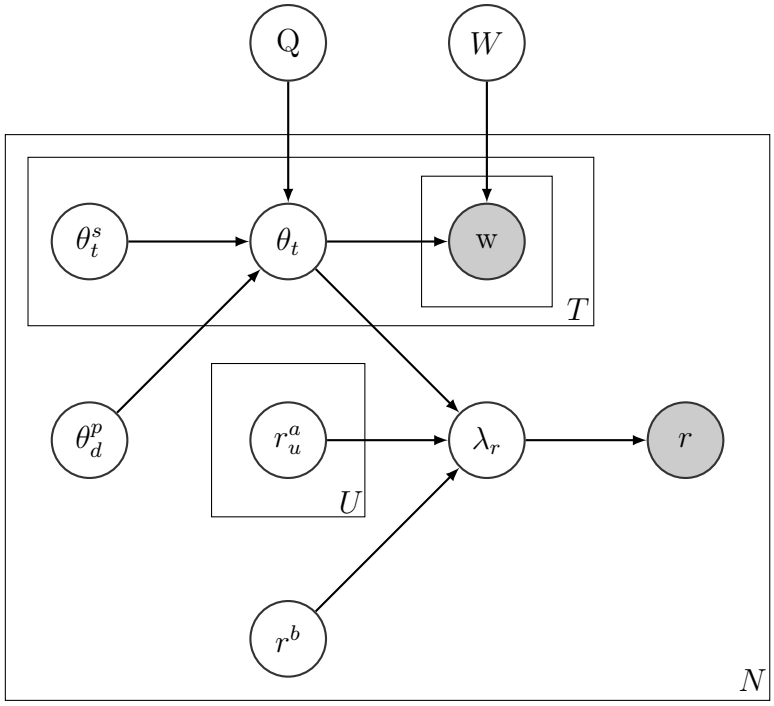

Figure 3.1: The graphical representation of the attribute-sentiment analysis sentence model

of the sentiment rating of the residual rating contributes to the overall rating in the sentence. We set $(\alpha = r, \beta = 1)$ with a mean of r as the prior for the sentiment ratings of the product attributes and the residual sentiment rating since these sentiment ratings should not deviate from the overall rating too much. Note there exists a trivial solution for r^b and r_u^a that all posteriors are set to be $(\alpha = 2r, \beta = 2)$ with a mean of r. In the discussion of the inference network in Section 3.3, we parameterize the parameters of the posteriors of the sentiment ratings of the product attributes as a function of the text content in the review. In theory, it is still possible to reach the trivial solution, but is unlikely to happen in practice.

As before, we use an approximation algorithm to compute approximated posteriors

of the sentiment ratings of product attributes and the residual sentiment rating. The detail of the derivation is Appendix B. We interpret the mean of the approximated posterior of the variables r_u^a and r^b as the ratings for product attributes and the residual rating. To ensure the correspondence to the rating scales, we enforce the function that parameterizes the mean parameter of the Gamma posteriors to be a sigmoid function with a scaling factor same as the rating scale. For an online review using a 5-star rating scale for the overall rating, we set the scaling factor to be 5.01. The sigmoid function and the scaling factor ensure the mean of the posterior is between 0 and 5.01. Given the Poisson regression function normalizes the weights to be 1, the mean of the posterior well corresponds to the rating scale.

Many existing multi-aspect sentiment analysis models (e.g. Jo and Oh, 2011; Moghaddam and Ester, 2011; Wang and Ester, 2014) rely on a set of sentiment seed words or the polarity score of a sentiment lexicon to correspond the sentiments of product attributes to sentiment ratings. They have trouble predicting sentiment ratings where the review system uses more than two rating scales. Even with the polarity information, the polarity score may not correspond well to the rating scales. Besides, such models leave out the readily available overall ratings that are informative. Our design models the overall rating as a weighted combination of the sentiment rating of product attributes through sentences.

Compared to the models (e.g. Wang et al., 2010, 2011; Wu and Ester, 2015) that utilize the overall rating to model the ratings of product attributes, we establish a clear correspondence between the sentiment ratings and the rating scale by parameterizing the parameters of the sentiment rating posteriors as a sigmoid function and a fixed scaling factor. As the parameters of the sentiment rating posteriors become a function

of the text content, the model is free to use the whole text content of the review to infer the sentiment ratings, different from other existing models that are limited to individual words and are unlikely to understand the sentiment of multi-words expressions such as "go/fall to pieces".

3.3 Inference Network

An inference network is a neural network that parameterizes the posterior distribution of a generative model enabled by backpropagation (Kingma and Welling, 2014; Rezende et al., 2014). The input of the inference network in the model is the text content of the review. The output is the parameters of the posterior distributions. The inference network has several advantages. Topic models often use hidden local variables, such as the inter and intra-sentence topics, the sentiment ratings of product attributes and the residual sentiment rating. During the training, both global variables such as the topic word matrix, and local variables of the training documents are learned. However, when applying the trained model to a new review, we need to infer the local variables. The inference usually involves an expensive optimization process: the inference of the attribute sentence model is less expensive due to the existence of a closed-form algorithm, while the attribute-sentiment analysis sentence model relies on the gradient descent algorithm using the approximated gradient and is much more expensive for inference. Using the inference network, the inference of local variables for a new review becomes a forward pass of the inference network with the text content of the review as the input, which is much faster than the optimization. The use of the inference network also improves the modelling of the sentiment

ratings of product attributes by using the full-text content to avoid a trivial solution and enforcing a clear correspondence to the rating scale.

The attribute-sentiment analysis sentence model has four types of local variables: the inter and intra-sentence topics, the sentiment rating of product attributes and the residual sentiment rating. One approach is to construct four independent neural networks for the posteriors of each type of local variable (Miao et al., 2016). However, the approach introduces too many parameters, increases the model complexity, makes the optimization harder, and ignores the relationship between different types of local variables. At the same time, this approach requires a neural network to learn the posteriors of the sentiment ratings of product attributes with only the gradient of the sentiment ratings and without the information of the product attributes; the learning would be impossible. Two recent studies (Howard and Ruder, 2018; Liu et al., 2018) demonstrate the capability of the neural network in transfer learning and multi-task learning. We examine the possibility of an inference network design with four outputs, each corresponding to one type of variables, using a shared word-embedding and bottom layers.

The challenge of designing such an inference network with the shared structure is that the posteriors of different types of variables require different information. The intra-sentence topic variables reflect the interaction of words in a sentence, the inter-sentence topic variables exhibit the interaction across sentences, but the information of word positioning is less useful for the two types of variables. The sentiment rating variables, different from the topic variables, are sensitive to the positioning information that is important for capturing negations and idiomatic expressions. Our design uses the convolutional neural network, instead of the standard multi-layer percep-

tron (MLP) neural network (Miao et al., 2016; Srivastava and Sutton, 2017), as the building block for the inference network to accommodate the different information needs of different types of variables. The convolutional neural network is essential in computer vision (e.g. Krizhevsky et al., 2012), and recently has been demonstrated to be effective in natural language processing (e.g. Gehring et al., 2017; Kim, 2014). Each convolutional block consists of two components: a convolution kernel and a rectified linear unit (RELU) non-linearity (Nair and Hinton, 2010). By stacking the convolutional block on top of each other, we design a neural network to model the interaction of words at both the sentence level and the document level.

From the input, we apply an intra-sentence stack of convolutional blocks to the word-embedding layer of the words in the sentence. Different sentences share the same stack of convolutional blocks. The input of the stack of blocks is of size $J \times |t|$ where J is the dimension of the word-embedding, and $|t|$ is the number of words in the sentence t and the output is of the same size. Each convolutional block has a convolution kernel of size $J \times J \times L^s$ and stride 1 where L^s is the width of the convolution kernel, such that the output of the convolutional block is compatible with the input of the block in the next layer. We pad both ends of the input to the convolutional block with $\mathbf{0}$ vectors to ensure the output is of the same size as the input, usually called "same padding" (Goodfellow et al., 2016). The structure of a convolutional block is illustrated in Figure 3.2. Note for a sentence of length 20, every output of the network depends on all inputs after stacking six convolutional blocks on top of each other, and each block uses a kernel of width 4. It allows capturing the interactions inside the sentence. The output of size $J \times |t|$ from the block goes through a max-pooling layer (Boureau et al., 2010), which chooses the maximum

value from the vector of size $|t|$ to form a vector of size J, a hidden representation of the sentence or the sentence embedding, denoted as h^s. We apply one fully-connected layer for the α parameter, and one fully-connected layer for the mean parameter to the hidden sentence representation; both fully-connected layers have the input of size J and the output of size K^s. The two output vectors go through a leaky RELU (Nair and Hinton, 2010) layer to compute the α and the mean ($\frac{\alpha}{\beta}$) parameters for the posteriors of the intra-sentence topic variables of sentences. The leaky RELU avoids the invalid negative and 0 values. The structure of the output layer for the topic variables is described in Figure 3.3.

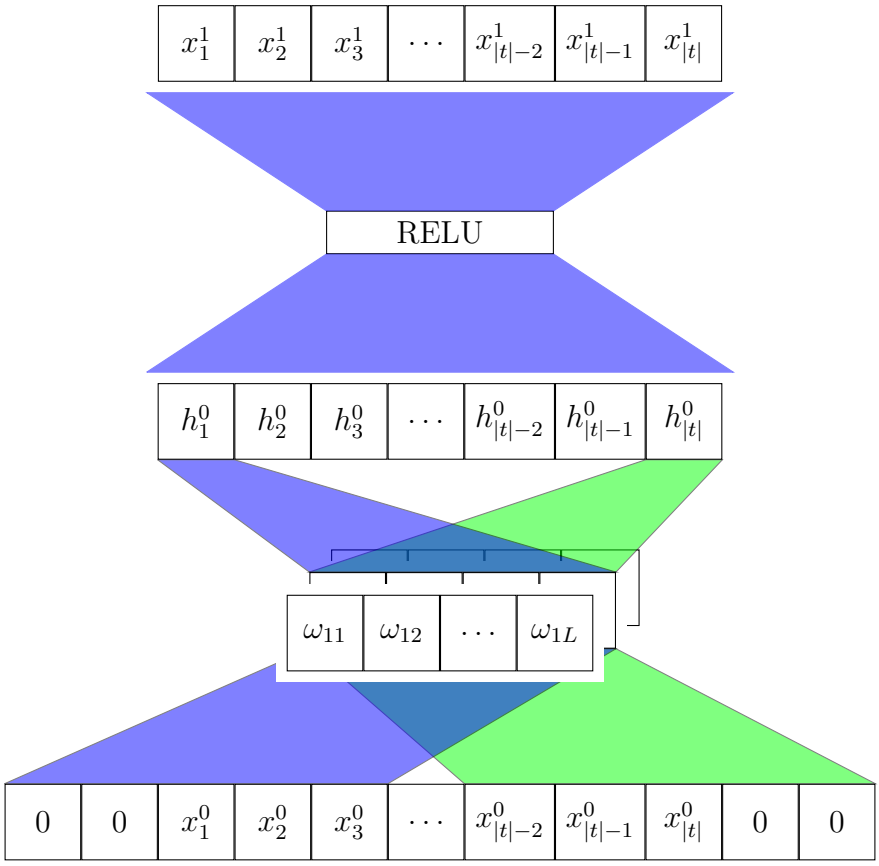

Figure 3.2: The structure of the convolutional block: each input cell x^0_i is a vector of size J; each convolution cell ω_{il} of a convolution kernel is a vector of size J; in total, there are J convolution kernels; the hidden cell h^0_i and the output cell x^1_i have the same size as the input cell.

To capture the interactions across sentences, we apply an inter-sentence stack of convolution blocks on the top of the concatenation of the hidden sentence representation of size $J \times T$. The convolution blocks in the stack use the kernel of the size $J \times J \times L^p$, similar to the blocks in the intra-sentence stack but using a different kernel width K^p. As before, the output from the block passes through a max-pooling to form a hidden representation of the review or the review embedding, denoted as h^p. Similarly, we compute the parameters for the inter-sentence topic posteriors of the review from the hidden representation by going through two separate fully-connected layers, both having the input of size J and the output of size K^p, and passing the leaky RELU layer. The network architecture uses two stacks of convolutional blocks to organize a review as a tree: the word embedding in the leaves, the sentence embedding in the middle and the review embedding at the root. The stack of convolutional blocks maps a sequence of word embeddings to a sentence embedding and then transforms a sequence of sentence embeddings to a review embedding. We extract the intra-sentence topic information from the sentence embedding and the inter-sentence topic information from the review embedding.

To infer the parameters of the sentiment ratings of product attributes and the residual sentiment rating, we introduce a sentiment embedding h^r that maps the overall rating to a continuous vector as in the word embedding. The sentiment embedding is superposed to the word embedding that the input to the inference network becomes $[x_1 + h^r, \ldots, x_n + h^r]$ where x_n is the word embedding of the word at position n. The inference network uses the additive attention mechanism to determine the dissemination of gradient across sentences (Bahdanau et al., 2014). We define the

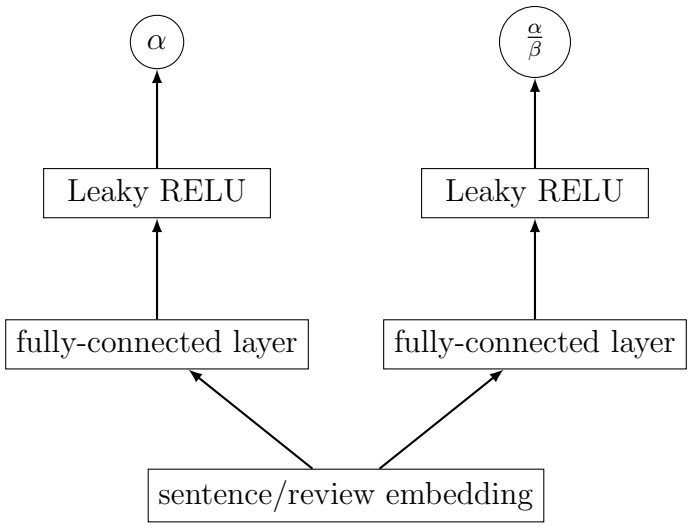

Figure 3.3: The output structure of topic variables

attention function for a ratable product attribute u in a sentence t as follows,

$$
\begin{aligned}
e_t^a &= v^\intercal \mathrm{RELU}(W_1^i h_t^s + W_2^i h^r) && (3.3)\\
f_{tu}^a &= \left(\left(\frac{\alpha_{\mathrm{Q}}^\gamma}{\beta_{\mathrm{Q}}^\gamma} \left(\frac{\alpha_{\theta_t^s}^\gamma}{\beta_{\theta_t^s}^\gamma} \right) \odot \frac{\alpha_{\theta_d^p}^\gamma}{\beta_{\theta_d^p}^\gamma} \right) \oplus \frac{\alpha_{\theta_t^s}^\gamma}{\beta_{\theta_t^s}^\gamma} \right)_{\mathrm{k}(u)} && (3.4)\\
\mathrm{Attention}_{tu}^a &= \frac{\exp(e_t^a) + f_{tu}^a}{\sum_{t'} \exp(e_{t'}^a) + f_{t'u}^a}
\end{aligned}
$$

The attention function for the residual sentiment rating in a sentence t is similar as follows,

$$
\begin{aligned}
f_t^b &= \sum_k (1 - I(\mathrm{k}(u) = k)) \left(\left(\frac{\alpha_{\mathrm{Q}}^\gamma}{\beta_{\mathrm{Q}}^\gamma} \left(\frac{\alpha_{\theta_t^s}^\gamma}{\beta_{\theta_t^s}^\gamma} \right) \odot \frac{\alpha_{\theta_d^p}^\gamma}{\beta_{\theta_d^p}^\gamma} \right) \oplus \frac{\alpha_{\theta_t^s}^\gamma}{\beta_{\theta_t^s}^\gamma} \right)_k && (3.5)\\
\mathrm{Attention}_t^b &= \frac{\exp(e_t^a) + f_t^b}{\sum_{t'} \exp(e_{t'u}^a) + f_{t'}^b}
\end{aligned}
$$

The attention function has the parameters v, W_1^i and W_2^i. The hidden representations for the sentiment rating of a product attribute u and the residual sentiment rating

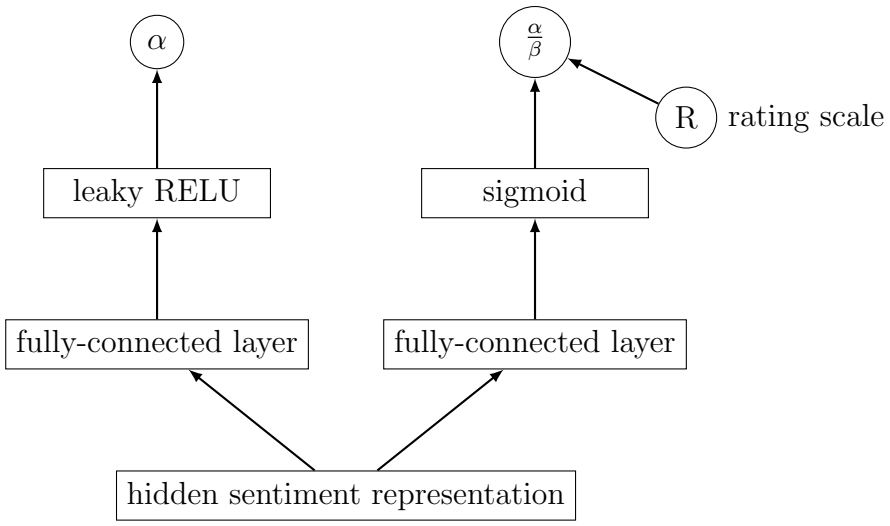

Figure 3.4: The output structure of attribute rating variables

are:

$$h_u^a = \sum_t \text{Attention}_{tu}^a \times h_t^s \tag{3.6}$$

$$h^b = \sum_t \text{Attention}_t^b \times h_t^s \tag{3.7}$$

As before, we apply two separate fully-connected layers: one followed by a leaky RELU layer to the hidden representation to extract the α parameter, the other followed by a sigmoid layer to extract the unscaled mean parameter. The scaling factor maps the unscaled mean parameter to the same scale as the rating scale of the review system. Both fully-connected layers have the input of size J and the output of size 1. The scaling factor is set to 5.01 for a 5-star scale rating. The structure of the output layer for the product attribute rating variable is described in Figure 3.4.

The attention functions for the sentiment ratings consist of two components: the e component (Eq. 3.3) and the f components (Eq. 3.4 and Eq. 3.5). The e component is consistent across the sentiment ratings of product attributes and the residual rating

and intends to reflect the similarity between the sentence and the overall rating. The idea is if the content of the sentence is consistent with the overall rating, for example, a sentence conveying the negative sentiment in a review with a low overall rating, the sentiment rating gradient should have more effect on the sentence than a sentence only stating facts. The f component replaces the random variables in Eq. 3.1 with their mean. If most words in the sentence are about the product attribute $\mathrm{k}(u)$, the sentiment rating gradient of the product attribute should affect the sentence more than a sentence unrelated to the product attribute. The residual sentiment rating gradient tends to influence the sentences that are unrelated to any ratable product attribute.

The inference network uses an intra-sentence stack of 6 convolutional blocks with a kernel of a width 4. Each output of this intra-sentence stack of blocks depends on 20 $(4 \times (6-1))$ input words from sentences. Since around 80% of sentences have less than 20 words in our datasets, this stack can model the interaction among all words of most sentences. We use an inter-sentence stack of 3 convolutional blocks with a kernel of a width 4 on top of the intra-sentence stack. This stack can model the interaction among 8 sentences, which fits well with our datasets that around 80% of reviews have less than 8 sentences. We apply dropout (Goodfellow et al., 2016) to the input of each convolutional block and use batch normalization (Goodfellow et al., 2016) before max pooling. The structure of the inference network is described in Figure 3.5.

Besides improving the inference speed of local variables, our inference network establishes an explicit link between the hidden sentiment rating variables of product attributes and the text content. It provides a new way of modelling sentiment in the

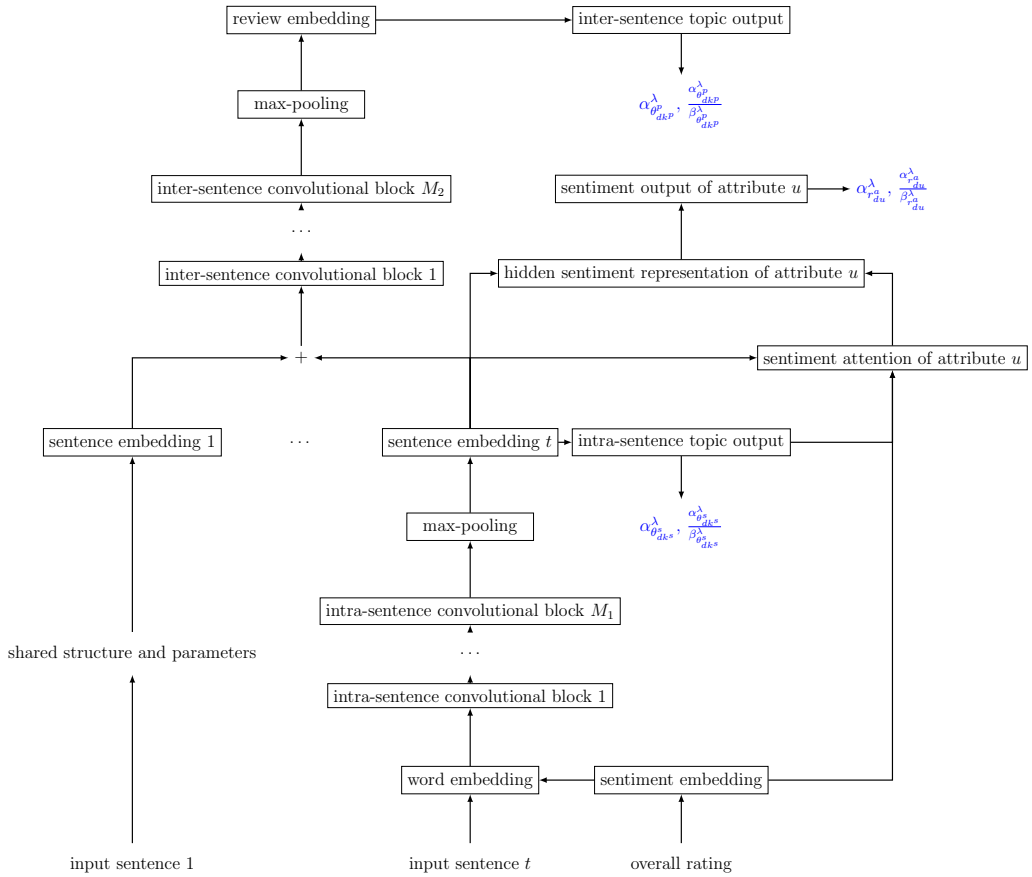

Figure 3.5: The structure of the inference network

topic model, different from the previous approaches (e.g. Jo and Oh, 2011; Wang and Ester, 2014; Wu and Ester, 2015), and greatly improves a topic model's capability in capturing sentiment. The design of our inference network that shares the structure among different types of local variables, uses the convolutional network and applies the attention mechanism effectively controls the number of parameters while maintaining a high modelling capacity.

3.4 Model Implementation

We use the same routine to pre-process the text data. For the training data, we apply the Stanford Core NLP library (Manning et al., 2014) to split the reviews into sentences and recover the word's lemma or canonical form. We remove the non-English reviews and the reviews with less than four sentences. We collect the tokens with five or more occurrences in the training data to develop the vocabulary for the inference network. We call this vocabulary the inference vocabulary. The inference vocabulary uses a unique "UNKNOWN" token to represent any word in the input to the inference network that does not exist in the vocabulary. From the vocabulary of the inference network, we further remove the stop words, punctuation, and use TF/IDF to select the top 10,000 words as the vocabulary for the attribute sentence model and the attribute-sentiment analysis sentence model. We call this vocabulary the model vocabulary. The attribute sentence model and the attribute-sentiment analysis sentence model only generate the word in the model vocabulary. If a review consists of words that are not in the model vocabulary, these words do not affect and are not involved in the computation of two models. The testing data also goes through the same process of splitting reviews to sentences and converting words to their canonical form. We use the inference vocabulary, and the model vocabulary developed from the training data to map the testing data to the input to the inference network and the words that the models generate.

Two different vocabularies are used because the computation of the topic models is much more expensive than that of the inference network, and it is helpful to control the number of words in the model vocabulary. Also, from the view of the topic

models, frequent words and stop words do not contribute to the understanding of the topics but affect the model performance in identifying coherent topics. Further, the inference network needs to detect negations and idiomatic expressions that convey sentiments; frequent words and stop words are often an essential component of such structures and expressions.

The implementation of the two models uses Python 3 and Tensorflow. The attribute sentence model uses the algorithm in Appendix A. To deal with a large amount of data (100,000 training and testing reviews), we use the stochastic variational inference (Hoffman et al., 2013) and set the step size as $(t+1)^{-0.7}$ and the batch size as 512 based on the training data used in the experiment. The priors of the global variables such as W and Q are set to $\alpha = 0.1$ and $\beta = 0.3$. The priors of the local variables θ_d^p are set to $\alpha_c^p = 0.1$ and $\beta_c^p = 0.1$. The choice of these parameters follows (Naesseth et al., 2017) that by setting α and α_c^p less than 1 encourages sparsity in topic word distributions and topic distributions. We set the local variables θ_t^s are set to $\alpha_c^s = 0.03$ and $\beta_c^s = 0.1$ to reflect that intra-sentence topics are more focused in sentences. We randomly initialize the parameters of the global variables as $\alpha = \log(1 + \exp(0.1 + 0.1a))$ and $\beta = \log(1 + \exp(0.3 + 0.1a))$ where a is the standard normal variable. When we need to initialize a topic with a certain keyword, we increase the α of the prior for the keyword by 10 and the β by 1. It is equivalent to assigning one review consisting of ten occurrences of the keyword to the topic. The prior ensures that the keyword and related words are much more likely to be assigned to the topic during the training, which allows the keyword to form the topic with relevant words.

The training of the attribute-sentiment analysis sentence model uses the stochastic gradient descent method and sets the step size the same as in (Naesseth et al., 2017).

Though training the model directly from a random initialization is possible, the speed of convergence is very slow, and the use of the inference network complicates the training. We use several methods to speed up the training process. First, we use the attribute sentence model to train the global variables W and Q. We apply the trained global variables from the attribute sentence model to the attribute-sentiment analysis sentence model. We fix the global variables, the sentiment embedding and the parameters of the fully-connected layers for extracting the parameters of the sentiment ratings, and train the inference network. Secondly, we use the pre-trained GloVe word-embedding of 100-dimension ($J = 100$)(Pennington et al., 2014) as the initialization for the word embedding during the training of the inference network. When the inference network converges, the inference network can well parameterize local variables. The optimization no longer fixes these variables and parameters and train all variables and parameters together until convergence. We use the mean of the posterior r_u^a as the predicted rating for a product attribute in a review.

3.5 Model Evaluation

3.5.1 Attribute Sentence Model

We use the Yelp dataset (Yelp, 2015) to evaluate the attribute sentence model (ASM) and choose the restaurant reviews from the dataset. The dataset consists of 630,550 reviews for 21,397 restaurants. The evaluation of the attribute sentence model focuses on its capability to extract consistent and explainable product attributes. In total, we use 50 topics and treat 30 of them as the inter-sentence topics and 20 of them as

the intra-sentence topics. The 60:40 split between the inter-sentence topics and the intra-sentence topics produces the maximum approximated likelihood $\log P(D|\alpha, \beta)$ of the training data among the splits from 10:90, 20:80, ..., to 90:10. We randomly select 100,000 reviews as the training data and another 100,000 reviews as the testing data. After training the attribute sentence model, we review the top 20 words of each topic to see if the identified topics are interpretable. Appendix C lists the top 20 words of a few selected topics. We use the perplexity to quantitatively measure how well the model can fit the data and the normalized point-wise mutual information (NPMI) to measure the topic coherence. The definition of the perplexity is;

$$\text{Perplexity} = \exp(-\frac{\sum_d \log P(d|W, \text{Q})}{\sum_d |d|})$$

where W and Q are the learned parameters from the 100,000 training reviews in training, and d is the other 100,000 testing reviews. The perplexity measures how well the learned model fits the unseen data. We can interpret perplexity as the number of choices for each word position in the review. Without prior information or models, the review randomly selects a word from the vocabulary for each word position. As a result, the number of choices is the size of the vocabulary. A model that fits the data better can utilize the topic information to inform the choice of words and thus reduce the number of choices and has a lower perplexity. The perplexity measure has been widely used in evaluating topic models (e.g. Blei et al., 2003; Hoffman et al., 2013).

The NPMI examines the co-occurrence of words from the same topic in a reference corpus and is showed to closely correlated to human judgement in evaluating the human-interpretability of identified topics (Lau et al., 2014). The definition of the

NPMI is,

$$\text{NPMI} = \sum_{k}^{K} \sum_{i=1}^{m-1} \sum_{j>i}^{m} \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)}$$

where K is the number of topics in the model, m is the number of the top words in the topic k, and w_i and w_j are the top words in the position i and j. The term $P(w_i)$ is the ratio of the documents consisting of w_i in all documents in a reference corpus, and the term $P(w_i, w_j)$ is the ratio of the documents consisting of both w_i and w_j in all documents. In the experiment, we use the English Wikipedia dataset consisting of 9,611,451 documents as the reference corpus and examine the top 20 words ($m = 20$).

We compare the attribute sentence model (ASM) against the standard LDA (Blei et al., 2003), the CTM (Blei and Lafferty, 2007), the NVLDA (Srivastava and Sutton, 2017), and the Sparse Gamma model (SGM) (Ranganath et al., 2015). The result is summarized in Table 3.2. Compared to SGM that also uses the sparse gamma prior, the attribute sentence model produces better perplexity and topic coherence due to effectively modelling reviews at the sentence level, separating the intra and inter-sentence topics and keeping the number of parameters in control. At the same time, the separation of the intra and inter-sentence topics offers great convenience in developing the personalization of online review in Chapter 4. Though the CTM produces better perplexity than ours, our model outperforms the CTM significantly in producing coherent topics. In the experiment, we notice that the models using normal distributions as the topic prior, for example, CTM and NVLDA, produce better perplexity, but the quality of the topics is no match to the models using Gamma-based prior (a Dirichlet distribution comes from Gamma distributions), either examined qualitatively or quantitatively.

Table 3.2: Perplexity and NPMI

	Perplexity (lower is better)	NPMI (higher is better)
LDA	1719	0.26
CTM	1453	0.13
NVLDA	1718	0.11
SGM	1652	0.28
ASM	**1617**	**0.30**

3.5.2 Attribute-Sentiment Analysis Sentence Model

The evaluation of the attribute-sentiment analysis sentence model (ASASM) uses the Tripadvisor dataset. The dataset consists of the ratings of 7 product attributes, "Value", "Rooms", "Location", "Cleanliness", "Business service", "Check in / front desk" and "Service", besides the overall ratings (TripAdvisor, 2015). After the standard data processing, we further remove the reviews that miss any product attribute rating. In total, the dataset consists of 111,019 reviews for 2,927 hotels. We randomly split the dataset in half as the training data and the testing data. To measure the model performance, we use the Root Mean Square Error (RMSE) defined as below,

$$\text{RMSE} = \left(\frac{\sum_d^D \sum_u^U (\hat{r}_u^a - \mu_{r_u^a})^2}{DK} \right)^{\frac{1}{2}}$$

where $\hat{r}_u^a$ is the ground truth of the sentiment rating of the product attribute u, and $\mu_{r_u^a}$ is the mean of the posterior of the sentiment rating of the product attribute as the predicted rating. We also use the Pearson correlation (Wang and Ester, 2014; Wang

et al., 2011; Wu and Ester, 2015) to measure how well the predicted product attribute ratings can preserve the relative order of product attributes in a review concerning their ground-truth ratings. The definition is,

$$\rho_A = \frac{1}{D}\sum_{d}^{D} \rho([\hat{r}_1^a, \ldots, \hat{r}_U^a], [\mu_{r_1^a}, \ldots, \mu_{r_u^a}])$$

where $\rho(\ldots)$ is the Pearson correlation function. During training and testing the model, the overall rating is known to the model but the product attribute ratings are unknown to the model.

The evaluation compares the model to a baseline that uses the overall rating as the prediction for all product attributes ratings. We compare the model to the two multi-aspect sentiment analysis (MASA) models: **ILDA** (Moghaddam and Ester, 2011) and **FLAME** (Wu and Ester, 2015). Note the implementation of the **ILDA** model uses the dependency grammar parser to parse sentences to pairs of aspect and opinion phrases, which provides improved performance than the original parsing algorithm (Wang and Ester, 2014). As the **ILDA** model does not have a definite correspondence between the sentiment topics and the rating scale, we use the mapping gives the lowest RMSE. The evaluation also includes the methods from the closely related aspect-based sentiment analysis (ABSA). The major difference between ABSA and MASA as discussed in Chapter 2 is the product attribute ratings in the training data are known to the aspect-based sentiment analysis but are unknown to the multi-aspect sentiment analysis. The state-of-the-art method in the aspect-based sentiment analysis tends to be a performance upper bound of the methods in the multi-aspect sentiment analysis as the aspect-based sentiment analysis has access to the product attribute rating information that is not available to the multi-aspect sentiment analysis. The first

Table 3.3: Product attribute rating predictions

	RMSE (lower is better)	Pearson correlation (higher is better)
Overall	0.888	0.151
SVM+BoW	0.403	0.761
SVM+GloVe	0.385	0.773
DMSCMC	0.349	0.852
ILDA	0.417	0.779
FLAME	0.401	0.831
ASASM	**0.366**	**0.851**

method extracts the uni-gram and bi-gram features from the review text and creates a support vector machine classifier for each product attribute. We call the method **SVM+BoW**. The second method uses the pre-trained GloVe word-embedding of 100-dimension ($J = 100$) (Pennington et al., 2014), instead of the uni-gram and bi-gram features, as the representation of the review text, and develops a support vector machine classifier for each product attribute. We call the method **SVM+GloVe**. The third method is a state-of-the-art neural network model, called **DMSCMC**, that uses the bi-directional long-short term memory and the hierarchical iterative attention model (Yin et al., 2017). The result is summarized in Table 3.3.

The result shows that our model is competitive in predicting product attribute ratings even when compared to the methods from ABSA in both measures. In general, the ABSA methods have an advantage over the MASA methods in terms of RMSE as

the ABSA methods have access to the product attribute ratings in the training data. Our model substantially outperforms **ILDA** and **FLAME** in RMSE as the topic models in **ILDA** and **FLAME** are unable to understand multi-words expressions conveying sentiments. For example, a sentence "When the conference was running, and 80 people were trying to turn up for breakfast between 7 and 8 am, everything went to pieces" in the review contributes to a 2-star rating for "Business service". Both **ILDA** and **FLAME** are unable to understand the negative connotation of "went to pieces" as "went" and "pieces" by themselves are common, and both words are not in a relatively top position of the negative sentiment vocabulary of the topic "Business service". Our model uses the pre-trained word embedding and a capable inference network structure that allows predicting a negative rating for the product attribute. Compared to **SVM+GloVe**, the performance gain of our model comes from the ability to distinguishing the general discussion and the discussion of different product attributes. The capability gives our model the advantage to make the correct predictions for the reviews that product attributes ratings vary significantly from the overall rating.

In terms of the Pearson correlation measure, both our model and **FLAME** outperform **SVM+BoW** and **SVM+GloVe** since these two ABSA methods treat the prediction of each product attribute rating as an independent task. **DMSCMC** partially addresses the problem through parameter sharing but does not model the interactions among product attribute ratings that limits its performance in the Pearson correlation measure. Our model captures the interactions among product attribute ratings through the product attribute rating prior and the regression function, which allows our model to outperform **FLAME** and closely matches **DMSCMC** despite

its superiority in RMSE.

3.6 Summary

In the chapter, we introduce a novel topic model to distinguish the topics across sentences and the local topics in the sentences that allows better modelling the text data and identifying interpretable topics. Based on the topic model, we integrate the sentiment analysis and utilize the inference network to strengthen its capability in capturing sentiments. Our evaluation shows that the model can identify high-quality topics that are easy to interpret and effectively predict product attribute ratings with performance closely matching the state-of-the-art method in aspect-based sentiment analysis. Our model is suitable for predicting the product attribute ratings in reviews where product attributes are not well-defined beforehand and product attribute ratings are not available. In Chapter 4, we use the attribute-sentiment sentiment analysis model to develop a personalization of online reviews to improve customers' decision-making.

Chapter 4

Personalization

Personalization of product offerings is a common and essential practice in e-commerce retailers. It helps customers quickly identify products of interest, allowing businesses to better meet the diverse needs of customers by improving customer decision making and reducing decision effort (Xiao and Benbasat, 2007). When making purchase decisions, customers increasingly rely on user-generated online reviews; some even consider the information in online reviews more credible and trustworthy than information provided by vendors (Bickart and Schindler, 2001). However, the amount of information in online reviews is often overwhelming and can prevent customers from extracting useful information for decision-making (Godes and Silva, 2012). To address the problem, many online retailers ask customers to vote on the "helpfulness" of a review. However, different reviews evaluate different attributes, and different customers assign different levels of importance to different attributes, such that a "helpful" review for one customer might not be helpful for another. We use an attribute-based personalization approach to address the following research question:

- How can we address the problem of customer and review diversity in the context of online reviews to recommend useful reviews based on customer preferences and improve product recommendation?

To address this question, we design a personalization approach for online reviews using the attribute-sentiment analysis model introduced in Chapter 3. We formulate the hypotheses of the personalization approach with consumer search theory (Stigler, 1961) and human information processing theory (Cowan, 1988).

4.1 Personalization of Online Reviews

In this section, we discuss the personalization approach from three perspectives: preference elicitation, preference matching and personalized product presentation. The approach proposed in this thesis contributes to preference matching and personalized product presentation.

4.1.1 Preference Elicitation

A customer faces many choices in the market. To make a choice, the customer must rank the options. Customer preference is revealed in the way the customer ranks the options (Samuelson, 1948). A personalization approach uses explicit and/or implicit preference elicitation methods to determine a customer's preference. Explicit preference elicitation often requires a customer to fill out a survey about preferences. The survey is usually designed based on product attributes, customer characteristics and needs. However, completing the survey requires significant cognitive effort from the

customer and could have a negative influence on the customer's adoption of the technology, user experience, and, consequently, impression of the product and business (Xiao and Benbasat, 2007). Further, the customer may not possess the knowledge of product attributes to specify preferences. Implicit preference elicitation examines a customer's behaviour, such as purchase and browsing history, or more general behaviours such as viewing time, the period when a customer examines a product or the description of a product (Parsons and Ralph, 2014). Implicit preference elicitation requires minimum effort from the customer and gradually learns the preference of the customer through the customer's behaviour. A drawback of this method is the learned preference of the customer suffers from the "pigeon-hole" effect, in which recommendations relate to products the customer examined before, and recommendations are limited to such products and lack diversity and novelty (Knijnenburg et al., 2012). As our personalization approach examines the customer's preference over product attributes, it can overcome the "pigeon-hole" effect. For example, suppose a customer is interested in a Japanese restaurant (product) and likes Sashimi/raw fish (product attribute). The personalization may recommend an Italian restaurant (product) that specializes in Carpaccio/raw fish and meat (product attribute).

We use explicit preference elicitation in the experiment since the implicit elicitation is not suitable for a time-limited experiment session. The explicit method allows participants to consciously evaluate the product in terms of the product attributes they choose. Later discussion presents a potential learning method of implicit elicitation to demonstrate the possibility of doing so, but it is not the focus of this research.

4.1.2 Preference Matching

In designing preference matching algorithms, two approaches have been widely used: collaborative filtering and attribute-based models (Ansari et al., 2000). The attribute-based model approach is popular as the result is easy to interpret, an advantage usually not enjoyed by the collaborative filtering approach. However, the prediction accuracy of attribute-based models falls behind collaborative filtering significantly, which may imply insufficient preference matching. We argue that the poorer performance of attribute-based models comes from a presumed and incomplete set of product attributes employed in the model, which does not capture the diversity among customers. We use the data-driven approach and incorporate existing knowledge of the product attributes of a product to extract a comprehensive set of product attributes from online reviews. Chapter 3 demonstrates that the comprehensive set of product attributes can accurately predict product attribute ratings. The result supports using an attribute-based model in personalization.

Preference matching matches products to customers and involves two tasks: modelling products and modelling customers. A unique component of this research is to model online reviews since the focus is to personalize them to improve customers' decision-making. Previous approaches (e.g. Fader and Hardie, 1996; Singh et al., 2005) model product attributes as binary variables to indicate whether a product possesses a product attribute. Instead of using binary variables, Parsons and Ralph (2014) use a pre-test to establish the relative importance of product attributes. Though, collaborative filtering also models the relative importance of "product attributes", it represents the product attributes as a combination of products that are hard to

interpret. Our approach models a product by examining the rankings of its product attributes from reviews. The product model can represent not only the relative importance but also the absence of a product attribute or the lack of discussion in a product. To model reviews, we take a view of review helpfulness (Mudambi and Schuff, 2010) and represent the review helpfulness through the discussion extensiveness of the product attributes in reviews measured by the "soft" word count. The "soft" word count means that a word may belong to two product attributes of a different portion. For example, the word "spicy" in "spicy salsa" may belong to two product attributes that 40% belongs to "Mexican food" and 60% belongs to "Food taste". We use the product model, the review model, and customer activities to infer customer preference in implicit preference elicitation.

4.1.2.1 Product Model

The product attribute rating of a product in the context of personalization is different from the product attribute rating prediction in Chapter 3. In product attribute rating prediction, a reviewer rates all product attributes regardless of whether the reviewer discusses the product attributes in the review. If the reviewer does not discuss a product attribute, it is reasonable to assume the reviewer rates the product attribute as the overall rating adjusted by the ratings of the discussed product attributes. For instance, suppose a review has a 4-star overall rating, but the discussed product attributes carry a negative sentiment. From the Poisson regression function in Eq. 3.3, the model tends to predict a 4-star or more rating for the product attributes that are absent from the discussion of the review. It is reasonable since the positive rated product attributes that are missing from the discussion may compensate

for the negative sentiment of product attributes in the discussion to reach a positive overall rating. This design works well for product attribute prediction. However, the personalization of online reviews is a different context. The personalization provides recommendations to customers based on reviews, and customers expect to find evidence that supports the recommendations in the reviews. The direct application of the attribute-sentiment analysis model (see Section 3.2) may recommend a product with an excellent overall rating though none of its reviews provides a discussion of the product attributes. To address the inconsistency, we use a different definition of product attribute rating for a product. In this definition, a review does not contribute to the product attribute rating if the product attribute is absent from the discussion. The formulation is as follows,

$$\begin{aligned} \tilde{\theta}_{dt\mathrm{k}(u)} &\underset{0\le k\le K}{\propto} \max\left(\theta_{dt\mathrm{k}(u)} - \frac{1}{K}\sum_{k'}\theta_{dtk'}, 0\right) \times \sum_{k'}\theta_{dtk'} \\ \tilde{N}_{pu} &= \mathrm{N}\left(\{d : d \in p, \sum_{t\in d}\tilde{\theta}_{dt\mathrm{k}(u)} > 0\}\right) \\ r_{pu} &= \frac{1}{\tilde{N}_{pu}}\sum_{d\in p}\mathrm{I}\left(\sum_{t\in d}\tilde{\theta}_{dt\mathrm{k}(u)} > 0\right) \times r^{a}_{du} \end{aligned} \tag{4.1}$$

The subscripts p denotes a product, d denotes a review, u denotes a ratable product attribute, K is the total number of product attributes, and t denotes a sentence. The function $\mathrm{k}(u)$ maps the ratable product attribute u to the topic k, and $\theta_{dt\mathrm{k}(u)}$ is the weight or the discussion extensiveness of the product attribute u in the sentence t. r_{du} is the product attribute rating of the review d, r_{pu} is the product attribute rating of the product p, and both have the range from 1 to 5. $\tilde{\theta}_{dt\mathrm{k}(u)}$ keeps the product attributes with the above-average weight, assigns 0 to the rest, and re-normalizes over all product attributes to the original scale. $\mathrm{N}(\cdot)$ counts the size of a set and $\mathrm{I}(\cdot)$

is the indicator function. $\tilde{N}_{pu}$ is the number of reviews that have at least one sentence that has the above-average number of words about the product attribute u.

The rationale for the term $\tilde{\theta}_{dt\mathrm{k}(u)}$ is that the product attribute rating r_{du} of a review contributes to the product attribute rating r_{pu} of a product only if the product attribute u is extensively discussed in the sentence t. If the product attribute is not extensively discussed anywhere in the review, we do not use the review to determine the product attribute rating of the product. Note that 0 in $\tilde{N}_{pu}$ and r_{pu} means the reviews of the product lack the evaluation of the product attribute. We may interpret 0 as the product attribute is absent from the product as no review of the product ever discusses the product attribute.

The product attribute rating of products ranks products in the market regarding product attributes. The ranking reflects the market position of the product regarding the product attribute. In personalization, it makes sense to recommend a product that outperforms the other products with respect to the product attribute, to a customer who considers the product attribute important. The attribute model AM_p of a product p is a vector where each component AM_{pu} of the vector represents the ranking of the product in the product attribute. The definition is as below,

$$AM_{pu} \underset{p}{\propto} \exp(r_{pu} - \lambda \frac{\sigma(r_{du})}{\tilde{N}_{pu}}) - 1 \tag{4.2}$$

$\sigma(r_{du})$ is the sample variance of the product attribute rating of the reviews that consist of extensive discussion of the product attribute. λ is a discounting factor. We set it to 0.01 in the experiment, as a pre-test on the dataset shows that the value removes products for which the standard deviation of the product attribute rating is more than 1 from the top recommendations. AM_{pu} is normalized to 1 over all the

products (thus using the proportional $\propto$ symbol) in the market to reflect the relative position of the product with respect to the product attribute.

The term r_{pu} in the definition gives a product higher rank in a product attribute if the sentiments in reviews of the product about the product attribute are positive, and a lower rank if the sentiments are negative. The terms $\sigma(r_{du})$ and $\tilde{N}_{pu}$ discount the product ranking in the product attribute if the sentiments in reviews are inconsistent or the product attribute rating of the product comes from a small number of reviews. The $\exp(\cdot)$ function spreads products with different ratings further apart to make the model follow Zipf's law. In the experiment, we require the recommended products to have $\tilde{N}_{pu} \geq 10$ from a pre-test on the dataset to ensure the reliability of the product attribute rating of a product and remove the products that lack the discussion of the product attribute. Note, before the normalization, AM_{pu} is 0 for a product lacks the discussion of a product attribute. The reason is that the product attribute rating of the product r_{pu} is 0, and the ratings of the reviews r_{du} are 0. The product has no effect on the rest of the products after the normalization. It reflects the situation that the market structure of a product attribute remains unchanged when introducing a product that does not possess the product attribute to the market. For example, the opening of a new Mexican restaurant only serves Mexican food does not affect the customers who are solely interested in Chinese food. The definition also embodies the relative importance of product attributes in a product. For example, suppose a restaurant is renowned for Chinese food but also serves Japanese food. Assuming the restaurant has a higher rank in Chinese food than in Japanese food, the value of Chinese food in the model tends to be larger than the value of Japanese food after normalization. It shows that Chinese food is more important than Japanese food for

the restaurant. Also, considering a case of a niche product, the product possesses a unique product attribute favoured by the niche market while the rest of the products do not own the product attribute. Consequently, the corresponding component of the product attribute is 0 for the rest of the products and some positive value for the niche product. After the normalization, the component of the niche product becomes the maximum value, 1, for the product attribute. It suggests the product attribute is the most important for the product and agrees with the market position of the niche product.

4.1.2.2 Review Model

We design the review model to reflect the perceived helpfulness of reviews. Previous research identifies various characteristics of reviews that determine their helpfulness, including review extremity, review length, linguistic style, and statement type (e.g., Mudambi and Schuff, 2010; Schindler and Bickart, 2012). However, none examines product attributes discussed in reviews. Our review model incorporates the previously identified factors such as review length and review extremity and uses the identified product attributes from review to expand one-dimensional helpfulness to multi-dimensional helpfulness according to product attributes. The review model AM_d, similar to the product model, is a vector. Each component AM_{du} of the vector represents the helpfulness of the review in the product attribute and is defined as below,

$$AM_{du} \underset{d}{\propto} \left(\sum_{t \in d} \tilde{\theta}_{dt\mathrm{k}(u)} \right)^{1-\lambda|r_{du}-r_{pu}|}$$

The term $\tilde{\theta}_{dtk(u)}$ represents the extensiveness of the product attribute u being discussed in the sentence t and the sum represents the extensiveness of discussion in the review. It comes from $\theta_{dtk(u)}$, which approximates the number of words used in the discussion of the product attribute since it is the parameter of the Poisson distribution that models the word counts in the review. λ is a discounting factor.

In the definition, the value of AM_{du} becomes bigger if the sum of $\tilde{\theta}_{dtk(u)}$ gets bigger. It conveys the idea that a review is helpful in informing the product attribute u if the review spends many words on discussing the product attribute. It materializes the prior research showing that review length positively contributes to helpfulness. On the other hand, the value of AM_{du} becomes smaller if the product attribute rating of the review is far from the product attribute rating of the product. It reflects the idea that the negative contribution of review extremity to helpfulness. We normalize the value over all the reviews in the product category. As a result, the shorter reviews are discouraged further while the lengthy discussion of product attributes is preferred. In the experiment, we set $\lambda = 0.01$ to penalize to the inconsistent ratings from a pre-test on the dataset.

4.1.2.3 Customer Model

The customer model describes customers' preferences over product attributes in choosing products. It is natural to use the discrete-choice model to represent a customer as a vector of measured attributes (McFadden, 1974), and the measured attributes are the value/utility the customer places over product attributes. However, one flaw of such a design is that the logit discrete-choice model allows unreasonable substitution patterns (Berry, 1994). Our design addresses the flaw by

differentiating the preferences over product attributes as two types. One consists of unsubstitutable preferences, meaning that a product attribute is a must to the customer, and the customer would only consider the product if the product possesses the required product attribute. For example, a customer might be interested only in hotels that have a swimming pool. The other is substitutable preferences, meaning that the utility brought by the product attribute can be substituted by the utility brought by other product attributes. For example, a customer might be interested in hotels that have a swimming pool but would consider a hotel that does not have a swimming pool if it provides other amenities suited to the customer's needs. Our recommendation algorithm treats the unsubstitutable preferences as constraints in optimizing the substitutable preference. The unsubstitutable preferences comprise a set of product attributes, denoted as AS_c. The substitutable preference, AM_c, is a vector as in the product model. Each component AM_{cu} is the weight of the product attributes, reflecting the way the customer ranks the product attributes. The value of AM_{cu} becomes bigger when the customer ranks the product attributes higher. We can formulate recommending products to a customer as an optimization problem,

$$\begin{aligned} \max_{p} \quad & U(c,p) = AM_c^{\mathrm{T}} AM_p \\ \text{subject to} \quad & \tilde{N}_{pu} > a, \quad u \in AS_c \end{aligned}$$

The subscript c denotes the customer. $U(c,p)$ is the utility function of the customer c for the product p. As discussed before, $\tilde{N}_{pu}$ is the number of reviews that contain extensive discussion of the product attribute u. A substantial number of such reviews of a product implies the product possesses the product attribute. If the number is small, the product may not possesses the product attribute. a is a threshold to safely

determine the existence of the product attribute in the product. In the experiment, we set $a = 20$ for reliable matching results.

We show how to develop the customer model in explicit and implicit preference elicitation. In explicit preference elicitation, customers can select the product attributes that are unsubstitutable, and rank the product attributes or give weights to the product attributes to indicate the relative importance. It is straightforward to code the information as the customer model AS_c and AM_c. However, implicit preference elicitation may be unable to uncover the unsubstitutable preference. The substitutable preference is obtainable by examining the customer's online activity. We give a formulation that computes the customer's substitutable preference based on what products the customer is interested in and which reviews the customer reads. It is as follows,

$$\mathrm{AM}_{c,k} = \frac{\sum_i \mathrm{AM}_{p_i u} + \sum_j \mathrm{AM}_{d_j u}}{\sum_k (\sum_i \mathrm{AM}_{p_i u} + \sum_j \mathrm{AM}_{d_j u})}$$

The subscriptions i and j represent the number of products in which the customer shows interest and the number of reviews the customer finds helpful, respectively. We assume a customer is interested in or purchases a product because the relative importance of the product attributes of the product matches the customer's preference; a customer finds a review helpful because the review extensively discusses the product attribute the customer ranks high. Note that the product model is normalized over all products, but the review model is normalized over all reviews. The two values in the product model and the review model are of different scales. The customer's interests in products and reviews contribute differently to customer preferences. The assumption is that the interest in a product provides more reliable information about

the customer's preferences than a review. In the experiment, we use the explicit preference elicitation since the method is more suitable for the situation.

4.1.3 Personalized Product Presentation

A personalized product presentation has two objectives: the first objective is to personalize the product offering based on the product attributes' ranking that matches the customer preference; the second objective is to assist the customer in making sense of the large volume of reviews and improve the customer decision-making. The recommendation algorithm aims to achieve the first objective. To accomplish the second objective, we introduce a personalized sorting design. Sorting is a widely used design feature in personalizations (Benlian, 2015; Tam and Ho, 2005). Past research shows that sorting can reduce the cognitive effort in decision-making (Häubl and Trifts, 2000). Examining online reviews requires substantial cognitive effort compared to examining product descriptions. Customers can benefit from a design that reduces cognitive effort. Further, the studies of personalizations find a personalization is more likely to be perceived useful if it reduces cognitive effort (e.g. Lee and Lee, 2009; Xiao and Benbasat, 2007).

Existing online review systems support sorting reviews by the time of writing, by the review ratings, or by the review helpfulness votes by customers. The sorting by the time of writing or the review rating is unlikely to organize the reviews by quality and relevance to customers. Customers need to actively search for information from pages of reviews, resulted in high search costs. In the process, customers are more likely to feel discouraged in finding the necessary information, stop searching prematurely

and end up with incomplete and poor quality information that compromises decision-making. Sorting by the review helpfulness votes presents the reviews in the order of the number of votes. Reviews with many votes are likely to consist of high-quality information but not necessarily relevant information to the customer. Without relevant information, customers have to continue searching and end up in the same situation as using sorting by time or ratings. The sorting by helpfulness votes is less effective in reducing cognitive effort and may compromise the decision-making of customers who are interested in niche product attributes. Also, it takes time to accumulate a sufficient number of helpfulness votes that are indicative of high-quality information. The existing "one-size-fits-all" sorting designs cannot accommodate customer diversity to assist customers effectively.

Our personalized sorting design uses the review and customer model to predict the expected helpfulness of a review to a customer and sorts the reviews in the order of the expected helpfulness of reviews. The expected helpfulness of a review is defined as below,

$$U(c, d) = AM_c^{\mathrm{T}} AM_d$$

Note that the expected helpfulness only involves the substitutable preference of the customer model, as the assumption is the recommended product already meets the unsubstitutable preference. The review rating only has a minimal effect on the expected helpfulness when the review rating strays from the average substantially. A review is expected to be helpful for a customer if the review extensively discusses a product attribute in which the customer is interested. The information conveyed by the review matching the information needs of the customer should determine the

sorting instead of the sentiments carried by the review. The sorting organizes reviews in the order of their quality and relevance to the customer, not sentiments. By placing high-quality and relevant reviews at top positions, the customer can acquire the necessary information with less cognitive effort and is more likely to reach a high-quality decision with confidence (O'Reilly, 1982).

4.2 Hypothesis Development

As discussed, our personalization design has two objectives. In this section, we use consumer search theory to analyze the design to understand how the design may achieve these objectives. Based on this analysis, we formulate our hypotheses.

4.2.1 Consumer Search Theory in Online Reviews

Consumer search theory accounts for the process of searching for products in the market and specifies the "optimal stopping point" when a customer should stop searching and select a product from the current offering to maximize utility (Stigler, 1961). The theory states the "optimal stopping point" is when the search cost surpasses the expected utility. In the context of this thesis, the customer searches for information about a product in online reviews. We assume the customer reads reviews sequentially, i.e., one by one in the order that the reviews are presented. After reading each review, the customer updates the expected utility of the product. For each customer, the expected utility of a product has an upper bound. The customer is likely to purchase the product when the expected utility approaches the upper bound. The search cost is the time and attention the customer commits to examining the reviews.

The optimal stopping in this context happens in two conditions: 1) the search cost of examining the next review surpasses the potential gain in the expected utility of reading the review, or 2) the accumulated search cost surpasses the expected utility of the product. The rationale for the first condition is that the potential gain from reading another review gradually diminishes, but the cognitive effort required to read and process information remains when the customer's expected utility approaches the upper bound. The extra search cost does not turn into additional gains in expected utility. Accordingly, the customer stops searching. At this point, the customer knows enough about the product and decides to make a purchase decision. For the second condition, the accumulated search cost keeps increasing as the search goes on. The accumulated search cost eventually surpasses the expected utility since the expected utility has an upper bound, but the search cost does not. When it happens, the product is no longer worth the additional effort of examining more of its reviews, and the customer stops searching. At the point of time, the customer is likely to make the purchase if the expected utility is close to the upper bound. Otherwise, the customer abandons the product.

4.2.2 Customer Preference on the Search Process

The above discussion establishes the point of time when a customer stops examining reviews and the likely decision of the customer regarding the product at the stopping time. Here, we explain how customer preferences over product attributes affect the expected utility of a product and the search cost. When modelling a product using an attribute model, the utility function of the product has three properties related to our

discussion (Samuelson, 1948). First, the utility function is bounded above. Second, the utility function is a monotonic increasing function of the value of the product attributes of the product. Third, the utility increase that comes from the increase of the value of a product attribute gradually diminish given the rest of the product attributes remain the same, i.e., "the law of diminishing returns" (Samuelson, 1948). In the context of online reviews, the increase of the value of a product attribute comes from reading reviews, which is associated with a search cost. Due to diminishing returns, after a certain point, the search cost surpasses the utility increase, i.e., reading more reviews to increase the product attribute is no longer worth the effort.

Customer preferences over product attributes describe how customers rank product attributes. For an increase of the same amount of value to the product attribute, a low ranking product attribute produces less expected utility than a high ranking product attribute for the customer. There exists a cap for each product attribute that the search cost, required to increase the current value of the product attribute to the cap, is equal to the utility increase from the product attribute increase. The search cost surpasses the utility increase when the value of the product attribute is increased beyond the cap. Assuming the search cost required to increase the value of any product attribute is the same, we can show that the cap of the high ranking product attribute is higher than the cap of the low ranking product attribute. The reason is that the utility increase, which comes from the high ranking product attribute increase to the cap of the low ranking product attribute, is bigger than the utility increase from the low ranking product attribute increase. The utility increase from the high ranking product attribute increase is also bigger than the search cost. Hence, the high ranking product attribute has more room to increase to reach its cap.

We discuss the effect of customer preferences on the expected utility through an example. In the example, there are two products. The first product possesses an outperforming product attribute A in which the customer shows no interest. The second product possesses an outperforming product attribute B, which the customer ranks highly. We assume the two products do not possess any other product attributes to simplify the discussion. For the first product, the customer reads reviews and spends a certain amount of search cost to uncover the value of the outperforming product attribute A. At a certain point, the uncovered value of the product attribute A reaches the cap and provides the customer with a certain amount of expected utility. For the second product, the customer spends the same amount of search cost and uncovers the same amount of value of the product attribute B. The same amount of uncovered value of the product attribute B produces more expected utility for the customer than the product attribute A as the product attribute B ranks higher. At this point, the uncovered value of the product attribute B has not reached the cap yet because a high ranking product attribute has a higher cap than a low ranking product attribute. When continuing the search, the net sum of the search cost to reveal more value of the product attribute A and the expected utility increase from the more revealed value of the product attribute A becomes negative. As a result, the negative net sum drains the expected utility that fuels the search and prevents the customer from uncovering more value of the product attribute A. When the search stops, the first product is unlikely to provide enough expected utility that reaches the customer's upper bound. The customer is likely to abandon the product. The customer is insensitive to the outperforming product attributes in which the customer has no interest and is unable to uncover more value of such product attributes. For

the second product, the expected utility of the second product is already higher than the first product at the time of continuing the search. Also, the cap of the product attribute B is higher and provides more room to support more search to reveal more value of the product attribute B, which in turn translates into more expected utility. At the time the customer stops the search, the expected utility of the second product is much bigger than the first product. The customer is more likely to purchase the product when the search stops.

Besides the influence on the expected utility of a product, customer preferences over product attributes lead to different search costs when the customer processes information in reviews. The theory of human information processing (Cowan, 1988) distinguishes two types of memory: working memory and long-term memory. The information stored in working memory is in a very accessible state, while the information stored in long-term memory requires extra effort to retrieve. Before a customer starts examining reviews of a product, the customer has a few product attributes in mind, which the customer tends to rank highly, and wants to examine these product attributes in the product. The customer has an efficient mental representation for these product attributes in working memory that is suitable for the task. When the customer reads a review discussing the product attributes, the customer can process the information in the mental representation efficiently with less cognitive effort and search cost. For example, if a customer is interested in Japanese restaurants, the customer expects to read reviews about sushi and sashimi and use the evaluation of sushi and sashimi in the reviews to update the expected utility of the restaurant with a little effort. However, when the customer comes across reviews about unexpected and unfamiliar product attributes, the customer has to retrieve the prior knowledge

of these product attributes from long-term memory and forms a new mental representation to accommodate these product attributes. Thus incurs substantial cognitive effort and search cost. These unexpected and unfamiliar product attributes have low rankings in the preferences of the customer. For example, if the customer comes across a review about risotto and lasagna when expecting a Japanese restaurant, the customer is likely to be confused and has to make extra effort to reconcile the evaluation of risotto and lasagna with a mental representation for evaluating Japanese restaurants.

4.2.3 Hypotheses

To help the discussion, we first categorize the content of reviews into three categories. The first category is low-quality content, which contributes little to the customer's expected utility but requires a certain search cost. Low-quality content often consists of brief and general positive or negative comments, such as "Love this place! Best food ever!", or lengthy background stories, such as when a review of a restaurant in Las Vegas discusses mostly the story of why and how the couple come to Las Vegas instead of the restaurant. The first kind of low-quality content may require a little search cost as the length of the content is short, but the second kind incurs a significant search cost. The second category is irrelevant high-quality content that consists of evaluations of low ranking product attributes. The irrelevant high-quality content contributes to the customer's expected utility moderately. However, because the caps of the low ranking product attributes are low, the increase of the expected utility becomes marginal after the customer examines a little such content; the search

cost of inspecting more such content overwhelms the potential gain in the expected utility. Furthermore, the search cost is also high for such content as the customer may be unfamiliar with the discussed product attributes. The customer is more likely to terminate the search process prematurely before more content can contribute to the expected utility. The third category is relevant high-quality content that consists of the evaluation of high ranking product attributes. The relevant high-quality content contributes substantially to the customer's expected utility, while it only requires a small search cost due to the customer's familiarity with the product attributes. As the cap of the high ranking product attributes is high, the expected utility can benefit from the customer examining more of such content. For each review the customer examines, the customer updates the beliefs in the product attributes discussed in the review. The belief is the customer's mental representation for evaluating the product attribute that stores the information about the product attribute. The customer increases the belief in the product attribute if the product attribute is favourably evaluated, otherwise, the customer decreases the belief. The customer estimates the expected utility and stops reading more reviews if any of the stopping conditions is met. The expected utility reaches the customer's upper bound, and the customer is likely to make a purchase decision, only after the customer examines enough relevant high-quality content. We summarize the customer decision process in the context of online review in Figure 4.1.

To examine the objectives of the personalized product presentation, we define three types of products. The first type is a recommended product. It is selected by the recommendation algorithm (see Eq. 4.3): it possesses the product attributes in which the customer is interested and has the best ranking in these product attributes.

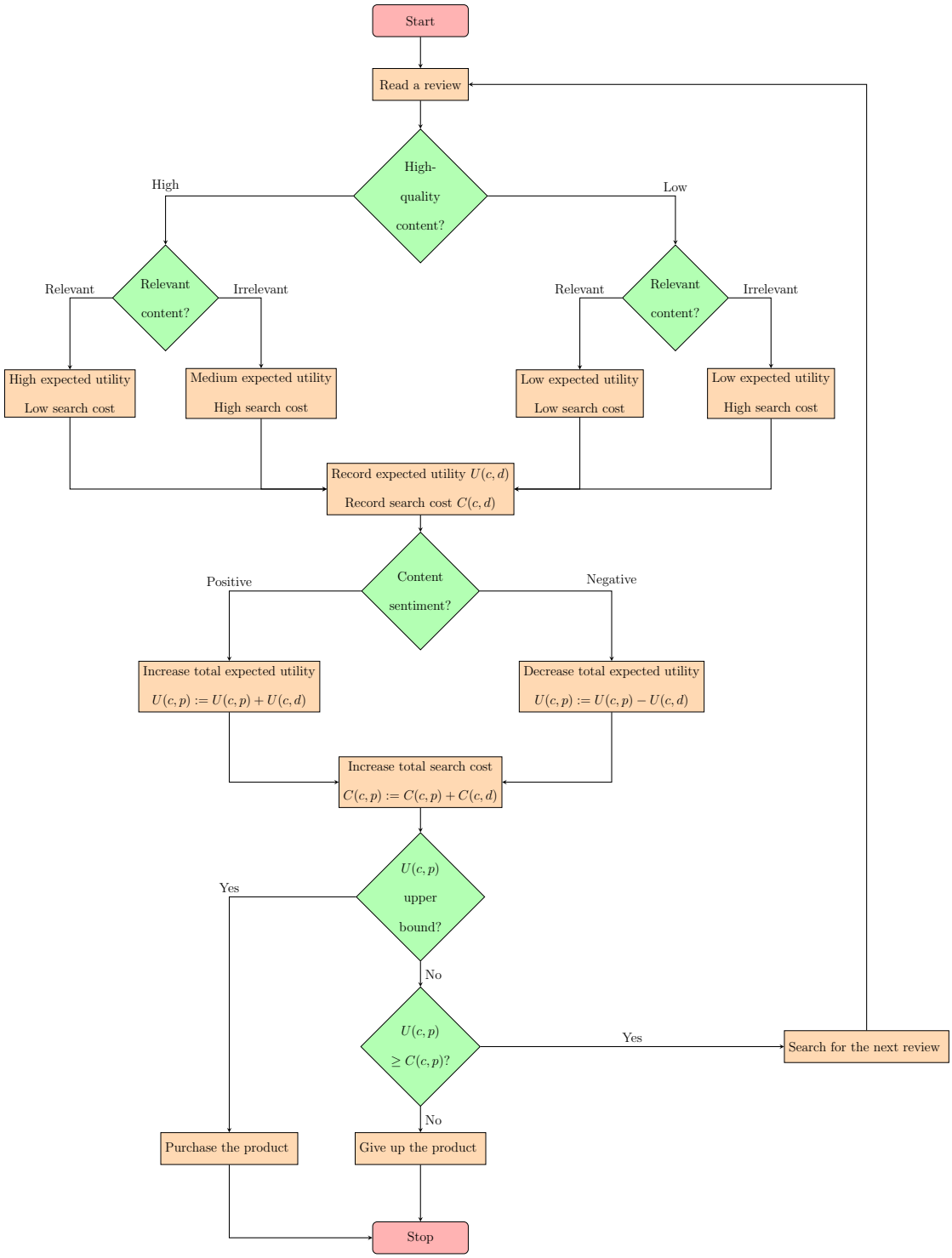

Figure 4.1: Customer decision process in the context of online reviews

The second type is a random product. It is randomly selected from all products and may not possess the product attributes. We also introduce a "bad recommendation," the third type, which possesses the product attributes, but has the worst ranking in these product attributes (minimizing $U(c, p)$ in Eq. 4.3 instead of maximizing).

4.2.3.1 Recommended Products

By setting up the problem as above, we examine how customers response to recommended products and random products. The recommended product has a high ranking in the product attributes in which the customer is interested. The high ranking of the product attributes of the recommended product comes from the product attribute ratings of the reviews, which indicates that the relevant content of the product attributes tends to be positive. The product model also ensures that a good number of reviews has an extensive discussion of the high ranking product attributes. The customer is likely to come across these reviews during the examination. The positive relevant content of these product attributes in these reviews increase the customer's belief in the product attributes and turn into more expected utility as the customer has a high cap for these product attributes. The expected utility of the recommended product has a good chance to reach the customer's upper bound and leads the customer to a purchase decision. A random product may not possess the necessary product attributes. The lack of the relevant content of the necessary product attributes prevents the expected utility of the random product from reaching the customer's upper bound and shortens the search process as the overall search cost can easily surpass the expected utility. Also, the random product tends to have more irrelevant content about product attributes not of interest to the customers. It

requires more effort from the customer to process the irrelevant content and drives up the search cost. The increase of the search cost further reduces the search process. The customer is more likely to spend less time and read fewer reviews of the random product, and abandon the random product at an early stage of examination. From the discussion, we have the first hypothesis as follows:

- **H1**: Customers will prefer recommended products over random products of the same overall rating.

4.2.3.2 Personalize Sorting Design

Beside recommending products, our personalization uses the personalized sorting design to organize reviews. We examine the personalize sorting design from three perspectives: behavioural intention towards products, decision effort and decision quality. We use the time sorting design and the vote sorting design as the two baselines for comparison: the time sorting design sorts reviews by the time of writing; the vote sorting design sorts reviews by the helpfulness/useful votes. These two are the most common sorting design used in practice. We assume the products being presented by different sorting designs have the same overall rating. The assumption mitigates the potential effect of the overall rating and allows the analysis to focus on the sorting designs. The implication of the assumption is that most reviews of the products tend to be positive as the recommended product usually has a positive overall rating, and the two other types of products have the same overall rating as the recommended product.

First, we examine the customer's intention towards products under different sorting designs. When reviews are sorted by the time of writing, the content of the

reviews at the top positions can be any content of the three categories. The low-quality content contributes little to the expected utility but demands a significant search cost. The irrelevant content contributes to the expected utility but a low extent. The increase of the expected utility from such content is unlikely to support comprehensive search. The customer stops the search when the overall search cost overwhelms the expected utility. When the search stops prematurely, the customer may not come across enough relevant high-quality content. The expected utility is unlikely to reach the customer's upper bound, and the customer is more likely to abandon the product regardless of the product type. We expect the customer to have a lower intention towards the recommended product using the time sorting design than using the personalized sorting design. The premature stopping of the search diminishes the difference in the expected utility among different product types. If the customer quits the search before examining more positive relevant content in the recommended product, the expected utility of the recommended product is unlikely to be high. If the customer abandons the search before examining the negative relevant high-quality content in the bad recommendation, the expected utility of the bad recommendation is unlikely to be low. As a result, we expect the intention difference between the recommended product and the bad recommendation is less using the time sorting design than using the personalized sorting design.

When using the helpfulness votes to sort reviews, the content of the reviews at the top positions is usually high-quality as the number of helpfulness votes is an indicator of quality. But the relevance of the reviews at the top positions depends on the product types. The personalization recommends a product because the product is superior at the product attributes the customer ranks high. As the recommended

product usually has a positive overall rating, the reviews often voted as helpful are likely to be positive. The positivity of the positive reviews is likely due to the superior product attributes, i.e., the helpful reviews are likely to discuss the product attributes the customer ranks high and evaluate them positively. For the recommended product, there is a good chance that many reviews in the top positions consist of relevant high-quality content. More of the relevant content combined with some irrelevant content pushes the expected utility over the customer's upper bound that leads to the purchase decision. We expect the customer has a greater intention towards the recommended product using the vote sorting design than using the time sorting design.

The story is a bit different for the bad recommendation and the random product. Note both the bad recommendation and the random product have the same overall rating as the recommended product. The bad recommendation, however, has a rather poor performance in the product attributes the customer ranks high. As before, the helpful reviews of a product with a positive overall rating are likely to be positive. The positive helpful reviews are less likely to discuss the poor-performed product attributes. Otherwise, they would be negative. As a result, the reviews in the top positions often consist of the irrelevant high-quality results, while the vote sorting design pushes the reviews consisting of relevant content to the back. Though the irrelevant content contributes to the expected utility, more of such content is unlikely to increase the expected utility substantially due to the low cap of the product attributes. In the meantime, such content incurs a high search cost preventing a comprehensive search. The bad recommendation faces a similar dilemma in the time sorting design. The expected utility is not high enough to reach the customer's upper bound leading to a purchase decision. At the same time, the expected utility is not

low enough that forces the customer to abandon the product early. We expect the intention difference between the recommendation and the bad recommendation to be less using the vote sorting design than using the personalized sorting design. The same analysis applies to random products as well.

The personalized sorting design places the reviews with the relevant content at the top positions. It increases the expected utility of the recommended product significantly since the relevant content of the recommended product is more likely to be positive. As the product attributes discussed in the relevant content have a high cap, more of the relevant content turns into more increases in expected utility. The customer encounters much positive relevant content of the recommendation when reading the reviews at the top positions. The positive relevant content pushes the expected utility over the customer's upper bound and leads to the purchase decision. We expect the customer has a greater intention towards the recommended product using the personalized sorting design than using the time sorting design. The personalized sorting design is also effective for the bad recommendation but in the opposite way. The expected utility of the bad recommendation decreases significantly from reading reviews in the top positions since the relevant content of the recommendation is more likely to be negative. As the expected utility drops substantially from reading the reviews in the top positions, the customer may find the expected utility of the bad recommendation is no longer worth continuing the search. Since the customer is more likely to expose to the negative relevant content in the reviews at the top positions, the expected utility at the time when the search stops tends to be low; the customer's intention to the product is low; the customer abandons the product. We expect the intention difference between the good recommendation and the bad recommendation

is greater using the personalized sorting design than other sorting designs.

By analyzing the types of reviews on the top positions in different sorting designs, we formulate the above discussion as the following hypotheses:

- **H2a**: The personalized sorting design will improve the customer's behavioural intention towards the recommended product over the time sorting design.
- **H2b**: The vote sorting design will improve the customer's behavioural intention towards the recommended product over the time sorting design.
- **H3**: The personalized sorting design will enable the customer to best distinguish the bad recommendation from the recommended product.

Next, we examine the decision effort required for different sorting designs. As discussed, the content of the reviews at the top positions in the time sorting design may be low-quality and irrelevant. As a result, the expected utility, on the one hand, is unlikely to increase to the upper bound for the recommended product. On the other hand, the expected utility is unlikely to decrease at a fast pace that accelerates the stopping for the bad recommendation and the random product. The reason is that most reviews tend to increase the expected utility as the overall rating is positive and the majority of them as positive. Both the negative low-quality content and the irrelevant high-quality content only reduce the expected utility to a lesser extent due to the customer's lack of interest. The customer has to spend a lot of effort in processing a large amount of less useful information until the search cost catches up with the expected utility or finally coming across the negative high-quality content. We expect the customer spends more effort using the time sorting design than using the personalized sorting design.

As before, we need to consider the product type when discussing the decision effort under the vote sorting design. The vote sorting design reduces the customer's effort when examining the recommended product. It is because the reviews at the top positions consist of more positive relevant content. Such content increases the expected utility significantly. The expected utility is more likely to reach the upper bound with fewer reviews, saving the customer from reading more less useful reviews. The customer knows enough about the product attributes that matter to the customer to make the purchase decision. However, the effect of reducing effort does not apply to the bad recommendation and the random product, and, on the contrary, the vote sorting design tends to increase the effort. As discussed, the top reviews when examining the bad recommendation and the random product consist of positive irrelevant high-quality content, which increases the expected utility, but the expected utility is unlikely to reach the customer's upper bound. The vote sorting design can bank more expected utility to prolong the search than the time sorting design. The reviews consisting of the negative relevant content are pushed back by the vote sorting design. The customer has more budget to lengthen the search to read through a lot of irrelevant content before reaching the negative relevant content that can quickly stop the search. The extra effort required for processing the bad recommendation and the random product is likely to exceed the reduced effort for the recommendation. We expect the customer spends more effort using the vote sorting design than using the personalized sorting design.

As the reviews consisting of the high-quality relevant content are at the top positions in the personalized sorting design, the customer collects enough information about the product from fewer reviews at the top positions for the decision-making

and avoids weeding through less useful and low-quality content that the personalized sorting design pushes back. The personalized sorting design reduces the customer's effort when examining the recommended product and the bad recommendation but in different ways. The high-quality relevant content of the recommended product quickly increases and pushes the expected utility over the customer's upper bound and leads to the purchase decision, while the same type of content of the bad recommendation decreases the expected utility quickly, that the expected utility can no longer support any additional search cost, and accelerates the stopping. However, the personalized sorting design has less effect on the random product as the random product may not possess the product attributes the customer requires. If a random product does not possess any required product attribute, the random product does not have any relevant content in the reviews. The customer's behaviour tends to be similar to the behaviour using the vote sorting design as both designs favour high-quality content. The customer is unlikely to purchase the product, has a low intention, and spends a bit more effort to come to a decision. But, if the random product possesses all or partial required product attributes, the customer's intention towards the random product is between the recommended product and the bad recommendation; the customer spends less effort to reach the decision. Overall, the customer spends less effort using the personalized sorting design because the design protects the customer from exploring irrelevant and low-quality content. We expect the personalized sorting design reduces the customer's effort in decision making. To summarize the above discussion, we can formulate the hypothesis as follows:

- **H4**: The personalized sorting design will reduce the customer's effort in decision-

making.

Finally, we examine the customer's decision quality under different sorting designs. As discussed before, the use of the time sorting design tends to stop the search prematurely. The premature stopping of the search prevents the customer from collecting high-quality information about the product. The customer lacks a full picture of the performance of the product and is likely to compromise the decision quality. Different from the time sorting design, the vote sorting design prioritizes the high-quality content. The customer feels better informed after examining the high-quality content and gains confidence in the decision. We expect the perceived decision quality is high using the vote sorting design. The personalized sorting design places the reviews with the high-quality relevant content at the top positions. It ensures the maximum exposure of the relevant content to the customer whenever available, provides critical information for decision making. The customer feels better informed and more confidence in decision-making. We expect the personalized sorting design improves the decision quality. Thus, we have the hypothesis as follows:

- **H5a**: The personalized sorting design will improve the customer's decision quality over the time sorting design.

- **H5b**: The vote sorting design will improve the customer's decision quality over the time sorting design.

We want to emphasize the personalize sorting design does not rely on the anchoring effect (Tversky and Kahneman, 1974) to manipulate the customer's intention. For every unit of effort that the customer puts into examining reviews, the personalized sorting design picks the most relevant and high-quality review that, from the

customer's view, best approximates the remaining unseen reviews. Image a complete information condition where the search cost is 0, and the customer has unlimited memory and processing power. The zero search cost allows the customer to examine all reviews before making the decision regardless of the sorting designs. The unlimited memory and processing power allow the customer to use all information available in the reviews to calculate the expected utility and form intention. We compare the expected utility of the complete information condition to the expected utility under different sorting designs of the standard information condition where the search cost is not 0 and the customer has limited memory and processing power. The expected utility difference between the complete information condition and the standard information condition tends to be smaller using the personalized sorting design than using the other two sorting designs, especially for the bad recommendation. The reason is that the personalized sorting design always places the content that influences the expected utility the most on the top, regardless of whether the influence is positive or negative. The content at the later position is neither relevant nor high-quality and barely affects the expected utility. The personalized sorting design can best approximate the expected utility in the complete information condition than the other two sorting designs as they do not have the same property.

4.3 Summary

In this chapter, we introduce a personalization design that consists of the attribute-based preference model for products, reviews, and customers. We analyze the process of examining online reviews of different types of products using different sorting

designs using consumer search theory. Based on the analysis, we formulate the hypothesis. In Chapter 5, we discuss the experiment design that empirically tests the hypotheses and interprets the results.

Chapter 5

Evaluation

This chapter presents the design and implementation of an experiment to evaluate the personalization approach proposed in Chapter 4. The empirical results from the evaluation support the utility of the design. We discuss the implication of the personalization and the limitation.

5.1 Experiment Design

We designed an experiment to test the hypotheses. As discussed, the experiment investigates the customer's behaviour intention towards three types of products of the same overall rating using three sorting designs. The requirement of the same overall rating eliminates the effect of the overall rating on customer behaviour. It ensures the effect on the customer's behaviour comes from the differences in product types and sorting designs. It is also meaningful in practice as a customer starts examining reviews only when the customer has trouble distinguishing two products from product descriptions, overall ratings and other prominent information cues.

5.1.1 Data Preparation

The product and review data used in the experiment are taken from the Yelp dataset (Yelp, 2015). We choose restaurants as the product category of interest because this category has a large number of products and reviews, and the evaluation of the product of the category does not require particular expertise. The data set consists of 21,397 restaurants and 630,550 reviews. The reviews in the dataset use the 5-star rating system for the overall ratings. Some of the reviews have the number of useful votes. We use the first 100000 reviews to train the attribute-sentiment analysis model and apply the trained model to compute $\theta_{dtk(u)}$ (the extensiveness of a product attribute discussed in a review) and r_{pu} (the product attribute rating of a review) for all reviews. From the trained model, we identify ten product attributes and group the five inter-sentence topics into the food type category and the five intra-sentence topics into the restaurant aspect category. The food type consists of five product attributes: "Mexican", "Breakfast/Brunch", "Italian", "Steakhouse", and "Japanese". The restaurant aspect also has five attributes: "Customer Service", "Wait Time", "Dining Environment", "Value/Price", and "Family Friendly". In the experiment, we treat the food type product attributes as the unsubstitutable attributes and the restaurant aspect product attributes as the substitutable attributes. We use $\theta_{dtk(u)}$ and r_{pu} to compute the product model AM_{pu} and the review model AM_{du}.

To ensure the reliability of the experiment, we remove the restaurants that have less than 20 reviews or an overall rating that is less than 4.0 from the dataset. The first removal criterion is because the experiment is to examine how customers evaluate a product based on reviews. A product that lacks a sufficient number of reviews

cannot provide the necessary information for decision-making and is not suitable for the experiment. The second removal criterion is due to the recommendation algorithm. If the personalization selects the recommended product without constraints, the recommended product often has a 5-star overall rating, which considerably limits the options of the bad recommendation. During our pre-test, we notice the difference between the recommended product and the bad recommendation is indiscernible from the reviews. However, when the overall rating is constrained to be less than 4, the pool of the bad recommendation becomes bigger, and the differences become apparent. In practice, a customer is unlikely to select a product with a 3.5-star or less overall rating over other products with a 4.5-star or 5-star rating. However, many customers may consider a product with a 4-star overall rating. In the experiment, the personalization first selects the bad recommendation, with at least a 4-star overall rating, and uses the overall rating of the bad recommendation to limit the selection of the recommended product and the random product. After filtering the restaurants, the final dataset consists of 3280 restaurants.

5.1.2 Operational Measures

Testing the hypotheses requires measuring several constructs. Hypothesis **H1** requires measuring the customer's interest in a product. Many studies (e.g. Xiao and Benbasat, 2007) show the customer self-reported intention reflects the customer's interest. Besides customer's intention, previous research shows viewing time is a good indicator for and correlated with the customer's interest (Parsons and Ralph, 2014). We also consider the number of words the customer reads as a measure of the cus-

tomer's interest, as it is correlated with viewing time; a customer tends to read more about a product if the customer is interested in the product. Hypothesis **H4** requires measuring the customer's decision effort. The decision effort is measured in the viewing time and the number of words the customer reads. We realize both the viewing time and the number of words the customer reads are used to measure both the customer's interest and the decision effort, which may seem conflicting, but it is not. If a customer is interested in the product, when using the same sorting design, the customer is more likely to spend more time examining and reading more about the product than a product in which the customer has no interest. In measuring the decision effort, viewing time measures the time spent in completing a task using different sorting designs, which is equivalent to decision time or task completion time. The number of reviews the customer reads is similar to the number of subtasks the customer completes for a task. As reviews vary in length, the review itself is not ideal for representing a subtask of equal workload, the number of words is a better approximation. Both decision time and the number of subtasks are often used to measure effort in the literature (e.g. Goodhue, 1995; Xiao and Benbasat, 2007). We measure the customer's intention with the purchase intention and the perceived quality and measures the decision quality with the perceived decision quality and decision confidence. We design a questionnaire using a five-point Likert-type scale for these constructs. The details of the questionnaire are provided in Table 5.1.

Constructs	Measures	Source
Perceived quality	What do you think of the overall quality of the restaurant?	Brady and Cronin Jr (2001)
	How likely will the restaurant provide you a good experience?	
	How likely will you feel good about what the restaurant provides to its customer?	
Purchase intention	Are you interested in going to the restaurant?	Baker and Churchill Jr (1977)
	Would you like to try the restaurant?	
	Would you like to visit the restaurant when you happen to see it?	
	Would you actively seek out the location of the restaurant in order to go to the restaurant?	
Perceived decision quality	My answers were based on the best available information.	Dooley and Fryxell (1999)
	My answers were made based on valid assumptions.	
	My answers helps me achieve my objectives.	
Decision confidence	I am confident that the decision made is indeed the best for me.	Häubl and Trifts (2000)
	I am certain that I have made the best choice for me.	
	I am positively sure that the decision made is really the best choice for me.	

Table 5.1: Construct measurement

5.1.3 Website Design

We developed a website for the experiment. The website first collects the participants' preferences, asks participants to evaluate three products, and answers a questionnaire for each product. When a participant logs into the website, the system generates a unique study id for the participant and randomly assigns the participant to one of the three sorting design groups. The sorting design group determines which sorting design the website uses to display reviews to the participant. Participants are not aware of the sorting design group to which they are assigned.

The website presents three products: the recommended product, the bad recommendation, and the random product, for the participant to evaluate. Before presenting the products, the website collects the customer's preferences by asking the participant to select one food type that the participant is interested in and one restaurant aspect that the participant cares most about. The experiment treats food type as the unsubstitutable product attribute and restaurant aspect as the substitutable product attribute. To generate the recommended product and the bad recommendation, we select the restaurants with $\tilde{N}_{pu_1} \geq 20$ (see Eq. 4.1 for the definition) where u_1 is the food type. For example, a restaurant is specialized in Japanese food only if a large number of reviews of the restaurant are about Japanese food. It helps to filter out the Chinese restaurants that may offer sushi occasionally. From the restaurants specialized in the food type, we select the restaurants with $\tilde{N}_{pu_2} \geq 10$ where u_2 is the restaurant aspect. It ensures that all the candidate restaurants have a reasonable number of reviews discussing the restaurant aspect the participant cares most about. Among these candidates, we first choose the restaurant with the lowest AM_{pu_2} as the

bad recommendation and choose as the recommended product the restaurant that has the highest AM_{pu_2}, with the same overall rating as the bad recommendation. The bad recommendation often has a 4-star rating and, in a few occasions, a 4.5-star rating but no 5-star rating. After deciding on the recommendation and the bad recommendation, we randomly select a restaurant, which has the same overall rating as them without filtering on $\tilde{N}_{pu_1}$ and $\tilde{N}_{pu_2}$, from all the restaurants.

After preparing the three restaurants, the website presents them to the participant in random order. The product page displays the name of the restaurant, the overall rating, and the reviews of the restaurant. If a participant is in the time sorting design group, reviews are in the order of time of writing (most recent first). If in the vote sorting design group, reviews are in descending order of the number of useful votes and in the reverse order of the time of writing when the reviews have the same number of useful votes. In the personalized sorting design group, the reviews are in the descending order of AM_{du_2}, where AM_{du_2} is the review model and u_2 is the restaurant aspect. We record the time when the product page is loaded and presented to the participant as the start time of the task. In the product page, we replace the overall rating with 4.5 regardless of the real overall rating of the restaurant to ensure the rating value does not influence participants' behaviour. The product page lists at most five reviews at one time on the page and allows the participant to navigate the reviews from page to page. When the participant navigates to the next/previous five reviews, we record the time the participant spends on the five reviews. Based on a pretest, we assume the participant read all five reviews and record the number of words of the five reviews as the number of words the participant reads if the participant spends more than 25 seconds on examining the five reviews not seen before. We

ask the participant to proceed to the questionnaire once the participant forms an opinion about the restaurant. We record the time when the participant leaves for the questionnaire as the end time of the task.

The questionnaire page consists of 13 questions as in Table 5.1: seven about the participant's intention and six about the decision quality. Besides collecting the answers from the participants, we use the questionnaire page for data quality control. For each questionnaire, we randomly select one question from the 13 questions, duplicate the question and place the duplicate away from the original question. The participant needs to answer 14 questions in each questionnaire with one question duplicated. At the same time, we record the time the participant used to complete the questionnaire. We expect the participant to give the same answer to the question and its duplicate and to spend no less than 15 seconds for each questionnaire. The screen-shots of the website is provided in Appendix D.

5.1.4 Participants and Data

We hosted the experiment on Amazon Mechanical Turk (AMT). AMT is a crowdsourcing marketplace that enables individuals and businesses to outsource their tasks to a distributed workforce. The task in our experiment is to evaluate three products using one of the sorting designs. We direct the workers who are interested in participating in the experiment to the website and provide each worker with a unique completion code once the worker completes the task. The worker submits the completion code to AMT to claim the reward. The use of AMT offers access to a representative and diverse population and strengthens the internal validity of the

research (Paolacci et al., 2010). One problem of AMT is that the workers tend to be less attentive than the test subjects in a lab environment with an experimenter. We employ a data quality check to weed out the low-quality data provided by the inattentive workers. The data quality check expects the participant spends at least 25 seconds to review one product and at least 15 seconds to answer one questionnaire. The participant is expected to give the same answer to the question that we intentionally duplicate twice. We offer 0.5 USD to the worker for completing the task and an extra 2.5 USD bonus if the worker's result passes the data quality check. In total, we recruit 357 participants, out of which the results from 173 participants pass the data quality check where 57 use the time sorting design, 58 use the vote sorting design and 58 use the personalized sorting design. The descriptive statistics of the experiment's data are provided in Table 5.2.

Sorting Design		Time	Vote	Personalized
# of participants	Count	57	58	58
Total time used (in seconds)	Mean	1005	941	637
	Min	83	84	85
	Max	3793	2720	1818
	St. Dev.	830	651	407
Time used in examining recommended products (in seconds)	Mean	390	417	280
	Min	29	27	28
	Max	1459	2632	935
	St. Dev.	350	449	188
Time used in examining bad recommendations (in seconds)	Mean	359	317	188
	Min	27	28	25
	Max	2477	1205	848
	St. Dev.	409	293	162
Time used in examining random products (in seconds)	Mean	257	208	169
	Min	25	27	26
	Max	1480	820	1185
	St. Dev.	308	178	191

Sorting Design		Time	Vote	Personalized
Total # of words read	Mean	9990	10457	7098
	Min	1221	2865	3176
	Max	31017	26878	10652
	St. Dev.	7188	5574	1856
# of words read of recommended products	Mean	3856	3782	2490
	Min	474	916	827
	Max	18721	13238	4481
	St. Dev.	3373	2592	885
# of words read of bad recommendations	Mean	3632	3734	2346
	Min	235	819	759
	Max	17279	11299	4601
	St. Dev.	3562	2304	834
# of words read of random products	Mean	2502	2942	2261
	Min	263	470	779
	Max	8050	7859	5375
	St. Dev.	1850	1847	1015
Intention towards recommended products (1-5 scale)	Mean	3.98	4.31	4.45
	Min	1.86	1.57	2.00
	Max	5.00	5.00	5.00
	St. Dev.	0.94	0.71	0.63
Intention towards bad recommendation (1-5 scale)	Mean	3.85	3.93	3.27
	Min	2.07	0.93	1.21
	Max	5.00	5.00	5.00
	St. Dev.	0.96	0.97	0.95
Intention towards random products (1-5 scale)	Mean	3.63	3.71	3.70
	Min	1.71	1.00	1.57
	Max	5.00	5.00	5.00
	St. Dev.	1.03	1.16	1.10
Decision quality towards recommended products (1-5 scale)	Mean	4.35	4.45	4.40
	Min	2.83	3.50	3.00
	Max	5.00	5.00	5.00
	St. Dev.	0.50	0.46	0.48
Decision quality towards bad recommendation (1-5 scale)	Mean	4.33	4.39	4.30
	Min	3.00	3.50	2.67
	Max	5.00	5.00	5.00
	St. Dev.	0.54	0.48	0.53

Sorting Design		Time	Vote	Personalized
Decision quality towards random products (1-5 scale)	Mean	4.34	4.25	4.31
	Min	2.50	2.83	2.33
	Max	5.00	5.00	5.00
	St. Dev.	0.55	0.61	0.59

Table 5.2: Experiment data

To ensure the measurement validity, we examine the correlations among the questions that measure the same construct. The results in Table F.1 and Table F.2 show that the questions are highly correlated ($\rho \geq 0.7$), indicating they indeed measure the constructs. We test normality for all variables to meet the assumption of later analyses. We notice that the time used, the number of words read, and their related variables are not normal, but in logarithm, they pass the normality test (see Table F.3). We transform the time used, the number of words read and their related variables to logarithm for the analysis. The variables in logarithm have "(in log)" next to their names. Besides the normality test, we test the equality of variances for the variables to satisfy the assumption of ANOVA (see Table F.4). The details of the examination are in Appendix F.

5.2 Data Analysis

5.2.1 Comparison of Customers' Interests between Recommended Products and Random Products

Table 5.3 shows the means of the time used, the number of words read and customers' intention. In the design of the experiment, all participants are presented with the

	Recommended Products μ_{p_0}	Random Products μ_{p_1}	$P > \|t\|$	$\mu_{p_0-p_1}$ 95% CI
N	173			
time used (in log)	5.521	4.896	**p<.001**	(0.48, 0.77)
# of words read (in log)	7.884	7.638	**p<.001**	(0.14, 0.35)
Customers' intention	4.25	3.68	**p<.001**	(0.37, 0.77)

Table 5.3: Comparison of customers' interests between recommended products and random products

recommended product and the random product. The design allows the within-subject comparison and uses the one sample t-test for better power than the two-sample test. We compare the within-subject differences in the measures between recommended products and random products and use one sample/paired t-test to examine the mean of the differences ($\mu_{p_0-p_1}$). The results in three measures are significant, and the 95% confidence intervals in three measures are in favour of recommended products. The evidence suggests that customers prefer recommended products over random products (supporting **H1**).

5.2.2 Comparison of Customers' Intention among Sorting Designs

We use ANOVA to examine the means of customers' intention for recommended products (see Table 5.4), and the result is significant. In the analysis of the 95% confidence interval, the personalized sorting design and the vote sorting design both are better than the time sorting design in improving customers' intention for recommended products (supporting **H2a** and **H2b**). However, the difference between the person-

Sorting design	Time μ_{s_1}	Vote μ_{s_2}	Personalized μ_{s_0}	$P > F$	$\mu_{s_0} - \mu_{s_1}$ 95% CI	$\mu_{s_0} - \mu_{s_2}$ 95% CI	$\mu_{s_2} - \mu_{s_1}$ 95% CI
N	57	58	58				
Intention for recommended products	3.98	4.31	4.45	**p<.005**	(0.18, 0.75)	(-0.15, 0.42)	(0.05, 0.61)
Intention for bad recommendations	3.85	3.93	3.27	**p<.005**	(0.23, 0.93)	(0.31, 1.01)	(-0.28, 0.43)
Intention for random products	3.63	3.71	3.7				

Table 5.4: Comparison of customers' intention among sorting designs

alized sorting design and the vote sorting design is insignificant. The experiment, to some extent, confirms helpful reviews of recommended products match customers' preferences as customers using both the vote and personalized sorting designs have similar intentions for recommended products. However, the helpful reviews of bad recommendations do not agree with customers' preferences; the customer using the vote sorting design is likely to have less chance to come across the negative discussion of the relevant product attributes. As a result, the intention for bad recommendations is higher using the vote or time sorting design than using the personalized sorting design.

The intention difference between recommended products and bad recommendations informs whether the sorting design is capable of assisting the customer in distinguishing recommended products from bad recommendations. We first use the one-sample t-test (within-subject comparison) to examine the mean of the intention difference for three sorting designs (see Table 5.5). The intention difference is insignificant using the time sorting design but is significant using the vote and personalized sorting design. We then use ANOVA to analyze the means of intention difference using three designs, and the result is significant. The confidence interval analysis shows

Sorting design	Time	Vote	Personalized	$P > F$	$\mu_{s_0} - \mu_{s_1}$	$\mu_{s_0} - \mu_{s_2}$
	μ_{s_1}	μ_{s_2}	μ_{s_0}		95% CI	95% CI
Intention difference	0.13	0.38	1.17	**p <.001**	(0.62, 1.47)	(0.37, 1.21)
p-value	0.46	**p<.01**	**p<.001**			
95% CI	(-0.22, 0.48)	(0.09, 0.67)	(0.90, 1.44)			

Table 5.5: Comparison of customers' intention differences between recommended products and bad recommendations among sorting designs

that using the personalized sorting design leads to the most intention difference. It suggests that customers using the time or vote sorting design have trouble detecting bad recommendations from recommended products; customers are likely to be confused, choose a bad product, and end up with a bad decision. The customers using the personalized sorting design are more likely to recognize the problems in bad recommendations and have an excellent chance to avoid such bad products (supporting **H3**).

5.2.3 Comparison of Customers' Decision Effort among Sorting Designs

Table 5.6 shows the results from the ANOVA test of the means of the time used and the number of words read. The results of the total time used and the total number of words read are significant. The confidence intervals show the personalized sorting design reduces the decision effort comparing to the other two sorting designs (supporting **H4**). When examining the decision effort for different products, we find the personalized sorting design is efficient in helping the customer to identify bad recommendations; both the time used and the number of words read are significant,

Sorting design	Time μ_{s_1}	Vote μ_{s_2}	Personalized μ_{s_0}	$P > F$	$\mu_{s_1} - \mu_{s_0}$ 95% CI	$\mu_{s_2} - \mu_{s_0}$ 95% CI
Total time used (in log)	6.571	6.565	6.234	**0.02**	(0.036, 0.639)	(0.031, 0.631)
Time used in examining recommended products (in log)	5.584	5.622	5.358	0.24	(-0.105, 0.557)	(-0.065, 0.593)
Time used in examining bad recommendations (in log)	5.333	5.254	4.897	**0.03**	(0.048, 0.823)	(0.013, 0.743)
Time used in examining random products (in log)	5.030	4.938	4.721	0.21	(-0.047, 0.666)	(-0.138, 0.572)
Total # of words read (in log)	8.897	9.121	8.832	**0.01**	(0.031, 0.288)	(0.066, 0.511)
# of words read of recommended products (in log)	7.859	8.0389	7.754	**p<.01**	(0.053, 0.362)	(0.028, 0.541)
# of words read of bad recommendations (in log)	7.712	8.053	7.697	**p<.01**	(0.061, 0.287)	(0.084, 0.627)
# of words read of random products (in log)	7.504	7.775	7.632	0.11	(-0.181, 0.124)	(-0.110, 0.393)

Table 5.6: Comparison of customers' decision effort among sorting designs

and their confidence intervals do not overlap. For random products, the personalized sorting design is not particularly efficient compared to the other two.

Sorting design	Time μ_{s_1}	Vote μ_{s_2}	Personalized μ_{s_0}	$P > F$	$\mu_{s_0} - \mu_{s_1}$ 95% CI	$\mu_{s_0} - \mu_{s_2}$ 95% CI	$\mu_{s_2} - \mu_{s_1}$ 95% CI
Decision quality for recommended products	4.35	4.45	4.4	0.55	(-0.13, 0.23)	(-0.23, 0.13)	(-0.08, 0.27)
Decision quality for bad recommendations	4.33	4.39	4.3	0.61	(-0.23, 0.15)	(-0.28, 0.1)	(-0.13, 0.25)
Decision quality for random products	4.34	4.25	4.31	0.72	(-0.24, 0.19)	(-0.16, 0.27)	(-0.30, 0.13)

Table 5.7: Comparison of customers' decision quality among sorting designs

5.2.4 Comparison of Customers' Decision Quality among Sorting Designs

Table 5.7 shows the results from the ANOVA test of the means of customers' decision quality. The results do not support customers' self-reported decision quality (**H5a** and **H5b**). All participants are confident with their decisions and think highly of their choices though some of the participants consider bad recommendations are as good as recommended products. Our results in decision quality echo the concerns of finding appropriate measures for decision quality (Aral et al., 2013).

5.3 Discussion and Limitations

We introduce a personalization method for online reviews that considers customer and review diversity to recommend products and organize reviews based on customer preferences. The personalization uses product attributes and sentiments extracted from the text content of online reviews to model products and customers and to develop the recommendation algorithm and the sorting design. Both our analysis and empirical evidence show that customers prefer recommended products over random

products. The personalized sorting design improves customers' behaviour intention towards recommended products and reduces the overall cognitive effort in decision-making.

One interesting finding in the research is the personalized sorting design best distinguishes recommended products from bad recommendations compared to the other two sorting designs. It accomplishes the objective by providing customers with high-quality and relevant content to improve customers' intention towards recommended products and to reduce customers' intention towards bad recommendations. We argue bad recommendations of the same overall rating are the worst options for customers due to the poor performance in the product attributes customers care most about. At the same time, they have a better chance than random products to confuse the customer into choosing it since it possesses the product attributes the customer demands. An online review presentation design needs to enable customers to distinguish two types of products.

Appendix E provides an example of a pair of recommended products and bad recommendations. In the example, both restaurants are Mexican restaurants matching the customer's preference for Mexican food and have an impressive 4.5 overall rating. From the top five reviews, both the vote and personalized sorting designs show the recommended restaurant offers the best value (Table E.2). However, for the bad recommendation, only the personalized sorting design (Table E.3) indicates the restaurant is overpriced, while the time sorting design (Table E.1) and the vote sorting design (Table E.3) show no such clue. It is impossible to know that the bad recommendation is overpriced without carefully read through many reviews unless the sorting design places the relevant reviews at top positions. When the relevant

content is misplaced, customers are much more likely to end up with bad recommendations. The vote sorting design can provide high-quality content to improve customers' intention towards recommended products, but the one-dimensional view of review content fails to reduce customers' intention towards bad recommendations. Its capability of differentiating recommended products and bad recommendations is limited; customers are at risk of ending up with poor choices.

We acknowledge that the research has several limitations. First, our personalization relies on the text content of reviews, disregarding other meta-information such as the time of writing and the authorship. Such information is vital in practice: a review from years ago is unlikely to be accurate, and a well-known reviewer or a friend is more credible than an anonymous reviewer. However, the limitation does not affect the results of the research. It is possible to extend the product and review models to integrate such information for better personalization. Secondly, our experiment does not indicate how reviews are sorted. For the vote sorting design, the numbers of helpful votes of reviews are not provided to the participants, which is different from practices where the numbers are shown to customers. Numbers serve as a signal for credibility and quality, thereby influencing the decision process (Chen et al., 2008). We may underestimate the intention towards recommended products and overestimate the decision effort for recommended products but not others since the most voted reviews of recommended products, not bad recommendations, match the customer's preference. The possible impact on the results is minimal.

Notwithstanding these limitations, our research demonstrates a novel design for personalization in a new context and a new analysis approach in understanding the effect of personalized online review.

Chapter 6

Conclusion

This research is inspired by the problem of customer and review diversity in the context of online review to recommend useful reviews based on customer preferences and improve product recommendation. The thesis answers the question with a personalization approach that uses attribute-based models to represent products, reviews and customer preferences.

As the attribute-based model using a pre-defined set of product attributes often falls short when predicting ratings, we explore the possibility of using a data-driven approach to identifying more comprehensive product attributes from online reviews to improve performance in sentiment analysis. In Chapter 3, we introduce a new topic model, the attribute-sentiment analysis sentence model, for extracting product attributes and predicting their sentiment ratings. The novel topic model considers word co-occurrences at the sentence level and the review level to identify different types of topics. This approach helps the model to extract more coherent topics than existing models and better fit the data than the models with the same Gamma family

priors. The use of an inference network with a shared structure for different posteriors enables a flexible parameterization of the posteriors and connects the sentiment ratings to the text content that improves the sentiment rating prediction performance.

Chapter 4 discusses and analyzes the design of the personalization of online reviews. We introduce the attribute-based representations of products, reviews and customer preferences in the context of examining online reviews. We develop a personalized product recommendation and review sorting design based on the representations to construct the personalization. Through consumer search theory and human information processing theory, we analyze the customer's behaviour of examining reviews given different types of products and using different sorting mechanisms and propose the hypotheses. We test the hypotheses with an experiment. The empirical evidence shows that: 1) the personalization of online reviews can recommend products matching customer's preferences; 2) the personalization can improve customer's intention towards recommended products; 3) the personalization can best distinguish recommended products from products that do not match customer's preferences; 4) the personalization can reduce decision effort.

The contribution of this research is multifold. First, it brings online reviews and personalization together to fill a gap in the area of the personalization of online reviews. Secondly, this research analyzes the process of examining online reviews of different types of products and using different sorting design. We base the analysis on consumer search theory and human information processing theory and verify the hypotheses with empirical evidence. Our work builds a theoretical foundation for analyzing the process of online reviews examination and enriches the theories.

Besides the contribution to the literature of personalization, the attribute-sentiment

analysis sentence model contributes to multi-aspect sentiment analysis. The model is suitable for review data, improves the coherence of the identified topics and allows them to be interpreted as product attributes. The use of the inference network addresses problems that topic models are less capable of understanding negative structures and multi-word idiomatic expressions, and often treat the sentiment analysis task as collaborative filtering. Our approach allows the model to achieve better sentiment rating prediction performance than the collaborative filtering approach and closely match the performance of the state-of-the-art aspect-based sentiment analysis method.

The results of this research suggest several directions for future study. First, we did not find support in this research that the personalization of online reviews improves decision quality. We suspect the failure in finding support may due to the use of the perceived decision quality to measure decision quality. We are interested in seeking a more objective measure of decision quality. At the same time, this search examines the effect of the personalization of online reviews on experience goods (restaurant). It would be interesting to examine how types of goods affect the decision process under the personalization of online reviews. Also, we only focus on user-generated online reviews and ignore expert reviews. In the future, we want to explore the integration and reconciliation between expert reviews and user-generated content in the models and the personalization design.

In this research, we propose a method for implicit preference elicitation but did not thoroughly examine the method. In the future, it would be valuable to examine the implicit preference elicitation method, identify the potential information cues that suggest customer preferences and derive customer preferences from such information

cues. Another direction for future research is to introduce new personalization designs. In this research, we present a personalized sorting design. In the future, we plan to examine the other personalization designs, such as information highlighting and review summarization. The personalization of online reviews provides a new research ground to test different theories.

Appendix A

Derivation of the Attribute Model

We introduce the intermediate variables $n^{k^p k^s}(w,t)$ and $n^{k^s}(w,t)$ where

$$\begin{aligned} n^{k^p k^s}(w,t) &\sim \text{Poisson}(W_{wk^p} \mathrm{Q}_{k^p k^s} \theta^s_{tk^s} \theta^p_{dk^p}) \\ n^{k^s}(w,t) &\sim \text{Poisson}(W_{wk^s} \theta^s_{tk^s}) \end{aligned}$$

and $n(w,t) = \sum_{k^p k^s} n^{k^p k^s}(w,t) + \sum_{k^s} n^{k^s}(w,t)$. We can interpret $n^{k^s}(w,t)$ as the number of word w from the intra-sentence topic k^s in the sentence t and $n^{k^p k^s}(w,t)$ as the number of word from the inter-sentence topic k^p adjusted by the intra-sentence topic k^s in the sentence. As each of intermediate variable is of a Poisson distribution, their sum is a Poisson distribution with the rate of the sum of their Poisson rate, consistent with the original definition. The posterior of intermediate variables is a multinomial distribution (Gopalan et al., 2015). The derivation of the attribute model

is as follows,

$$
\begin{aligned}
\log P(D|\alpha,\beta) &= \log \int d\,W\, d\,\mathrm{Q} P(W|\alpha_w,\beta_w) P(\mathrm{Q}|\alpha_c,\beta_c) \\
&\quad \times \prod_{d\in D} \int d\,\theta_d^p\, P(\theta_d^p|\alpha_c^p,\beta_c^p) \\
&\quad \times \prod_{t\in d} \int d\,\theta_t^s\, P(\theta_t^s|\alpha_c^s,\beta_c^s) \prod_{w\in t} P\Big(n(w,t)|W_w^\intercal\Big((\mathrm{Q}(\theta_t^s)\odot\theta_d^p)\oplus\theta_t^s\Big)\Big) \\
&\geq \mathbb{E}_{q(W)}\Big[\log P(W|\alpha_W,\beta_W) - \log q(W)\Big] \\
&\quad + \mathbb{E}_{q(\mathrm{Q})}\Big[\log P(\mathrm{Q}|\alpha_c,\beta_c) - \log q(\mathrm{Q})\Big] \\
&\quad + \sum_{d\in D} \mathbb{E}_{q(\theta_d^p)}\Big[\log P(\theta_d^p|\alpha_c^p,\beta_c^p) - \log q(\theta_d^p)\Big] \\
&\quad + \sum_{t\in d, d\in D} \mathbb{E}_{q(\theta_t^s)}\Big[\log P(\theta_t^s|\alpha_c^p,\beta_c^p) - \log q(\theta_t^s)\Big] \\
&\quad + \sum_{w\in t, t\in d, d\in D} \mathbb{E}_{q(\cdots)q(n_{wt})}\left[\log P\Big(n^{k^pk^s}(w,t)|W_{wk^p}\mathrm{Q}_{k^pk^s}\theta_{tk^s}^s\theta_{dk^p}^p\Big)\right] \\
&\quad + \sum_{w\in t, t\in d, d\in D} \mathbb{E}_{q(\cdots)q(n_{wt})}\left[\log P\Big(n^{k^s}(w,t)|W_{wk^s}\theta_{tk^s}^s\Big)\right]
\end{aligned}
$$

where the inequality is from Jensen's inequality and the terms after the inequality are the evidence lower bound (ELBO). We have the definition of $q(\cdots)$ as follows,

$$
q(\cdots) = q(W)q(Q)\prod_k q(\theta_{dk}^p)\prod_{kt} q(\theta_{tk}^s)
$$

By maximizing the ELBO, we can derive the closed form equations for the parameters of the variational distributions $q(\cdot)$ and use α^λ and β^λ to denote the variational parameters of the Gamma distribution. The update equation for $q(n_{wt})$ is,

$$
\begin{aligned}
q(n_{wt}=n^{k^pk^s}) &\propto \exp\Bigg((\psi(\alpha_{W_{wk^p}}^\lambda) - \log\beta_{W_{wk^p}}^\lambda) + (\psi(\alpha_{\mathrm{Q}_{k^pk^s}}^\lambda) - \log\beta_{\mathrm{Q}_{k^pk^s}}^\lambda) \\
&\qquad + (\psi(\alpha_{\theta_{tk^s}^s}^\lambda) - \log\beta_{\theta_{tk^s}^s}^\lambda) + (\psi(\alpha_{\theta_{dk^p}^p}^\lambda) - \log\beta_{\theta_{dk^p}^p}^\lambda)\Bigg) \quad \text{(A.1)} \\
q(n_{wt}=n^{k^s}) &\propto \exp\Bigg((\psi(\alpha_{W_{wk^s}}^\lambda) - \log\beta_{W_{wk^s}}^\lambda) + (\psi(\alpha_{\theta_{tk^s}^s}^\lambda) - \log\beta_{\theta_{tk^s}^s}^\lambda)\Bigg) \quad \text{(A.2)}
\end{aligned}
$$

The update equation for $q(\theta^s_{tk^s})$ is,

$$\alpha^{\lambda}_{\theta^s_{tk^s}} = \alpha_c + \sum_{w \in t} (\sum_{k^p} q(n_{wt} = n^{k^p k^s}) + q(n_{wt} = n^{k^s})) n(w,t) \tag{A.3}$$

$$\beta^{\lambda}_{\theta^s_{tk^s}} = \beta_c + \sum_{w \in t} \sum_{k^p} \frac{\alpha^{\lambda}_{\theta^p_{dk^p}}}{\beta^{\lambda}_{\theta^p_{dk^p}}} \frac{\alpha^{\lambda}_{W_{wk^p}}}{\beta^{\lambda}_{W_{wk^p}}} \frac{\alpha^{\lambda}_{\mathrm{Q}_{k^p k^s}}}{\beta^{\lambda}_{\mathrm{Q}_{k^p k^s}}} + \sum_{w \in t} \frac{\alpha^{\lambda}_{W_{wk^p}}}{\beta^{\lambda}_{W_{wk^p}}} \tag{A.4}$$

The update equation for $q(\theta^p_{dk^p})$ is,

$$\alpha^{\lambda}_{\theta^p_{dk^p}} = \alpha_c + \sum_{w \in t, t \in d} \sum_{k^s} q(n_{wt} = n^{k^p k^s}) n(w,t) \tag{A.5}$$

$$\beta^{\lambda}_{\theta^p_{dk^p}} = \beta_c + \sum_{w \in t, t \in d} \frac{\alpha^{\lambda}_{\theta^s_{tk^s}}}{\beta^{\lambda}_{\theta^s_{tk^s}}} \frac{\alpha^{\lambda}_{W_{wk^p}}}{\beta^{\lambda}_{W_{wk^p}}} \frac{\alpha^{\lambda}_{\mathrm{Q}_{k^p k^s}}}{\beta^{\lambda}_{\mathrm{Q}_{k^p k^s}}} \tag{A.6}$$

The update equation for $q(\mathrm{Q}_{k^p k^s})$ is,

$$\alpha^{\lambda}_{\mathrm{Q}_{k^p k^s}} = \alpha_c + \sum_{w \in t, t \in d, d \in D} q(n_{wt} = n^{k^p k^s}) n(w,t) \tag{A.7}$$

$$\beta^{\lambda}_{\mathrm{Q}_{k^p k^s}} = \beta_c + \sum_{w \in t, t \in d, d \in D} \frac{\alpha^{\lambda}_{\theta^s_{tk^s}}}{\beta^{\lambda}_{\theta^s_{tk^s}}} \frac{\alpha^{\lambda}_{\theta^p_{dk^p}}}{\beta^{\lambda}_{\theta^p_{dk^p}}} \frac{\alpha^{\lambda}_{W_{wk^p}}}{\beta^{\lambda}_{W_{wk^p}}} \tag{A.8}$$

The update equation for $q(W_{wk})$ is,

$$\alpha^{\lambda}_{W_{wk^p}} = \alpha_w + \sum_{t \in d, d \in D} \sum_{k^s} q(n_{wt} = n^{k^p k^s}) n(w,t) \tag{A.9}$$

$$\beta^{\lambda}_{W_{wk^p}} = \beta_w + \sum_{t \in d, d \in D} \sum_{k^s} \frac{\alpha^{\lambda}_{\theta^s_{tk^s}}}{\beta^{\lambda}_{\theta^s_{tk^s}}} \frac{\alpha^{\lambda}_{\theta^p_{dk^p}}}{\beta^{\lambda}_{\theta^p_{dk^p}}} \frac{\alpha^{\lambda}_{\mathrm{Q}_{k^p k^s}}}{\beta^{\lambda}_{\mathrm{Q}_{k^p k^s}}} \tag{A.10}$$

$$\alpha^{\lambda}_{W_{wk^s}} = \alpha_w + \sum_{t \in d, d \in D} \sum_{k^p} (q(n_{wt} = n^{k^p k^s}) + q(n_{wt} = n^{k^s})) n(w,t) \tag{A.11}$$

$$\beta^{\lambda}_{W_{wk^s}} = \beta_w + \sum_{t \in d, d \in D} \left(\sum_{k^p} \frac{\alpha^{\lambda}_{\theta^s_{tk^s}}}{\beta^{\lambda}_{\theta^s_{tk^s}}} \frac{\alpha^{\lambda}_{\theta^p_{dk^p}}}{\beta^{\lambda}_{\theta^p_{dk^p}}} \frac{\alpha^{\lambda}_{\mathrm{Q}_{k^p k^s}}}{\beta^{\lambda}_{\mathrm{Q}_{k^p k^s}}} + \frac{\alpha^{\lambda}_{\theta^s_{tk^s}}}{\beta^{\lambda}_{\theta^s_{tk^s}}} \right) \tag{A.12}$$

The algorithm iterates through all the update equations to compute the variational parameters.

Appendix B

Derivation of the Attribute-Sentiment Analysis Model

The derivation of the attribute-sentiment analysis model is similar to the derivation of the attribute model in Appendix A that relies on Jensen's inequality and a completely factorizable variational distribution. The difference is how we parameterize the local variational distributions $q(\theta_d^p|\alpha_{\theta_d^p}^\lambda, \beta_{\theta_d^p}^\lambda)$, $q(\theta_t^s|\alpha_{\theta_t^s}^\lambda, \beta_{\theta_t^s}^\lambda)$, $q(r_d^b|\alpha_{r_d^b}^\lambda, \beta_{r_d^b}^\lambda)$, and $q(r_{du}^a|\alpha_{r_{ud}^a}^\lambda, \beta_{r_{ud}^a}^\lambda)$ with an inference network that $(\alpha_{\theta_d^p}^\lambda, \beta_{\theta_d^p}^\lambda, \alpha_{\theta_t^s}^\lambda, \beta_{\theta_t^s}^\lambda, \alpha_{r_d^b}^\lambda, \beta_{r_d^b}^\lambda, \alpha_{r_{ud}^a}^\lambda, \beta_{r_{ud}^a}^\lambda) =$

$\mathrm{IN}(d, r)$ where $\mathrm{IN}(d, r)$ denotes the inference network.

$$
\begin{aligned}
\log P(D, R|\alpha, \beta) &= \log \int d\,W d\,\mathrm{Q} P(W|\alpha_w, \beta_w) P(\mathrm{Q}|\alpha_c, \beta_c) \\
&\quad \times \prod_{d \in D} \int d\,\theta_d^p\, P(\theta^p|\alpha_c, \beta_c) \\
&\quad \times \prod_{t \in d} \int d\,\theta_t^s\, P(\theta_t^s|\alpha_c, \beta_c) \prod_{w \in t} P\Big(n(w,t)|W_w^\intercal \theta_t\Big) \\
&\quad \times \int d\,r_d^b P(r_D^b|r, 1) \prod_u \int d\,r_{du}^a P(r_{du}^a|r, 1) \\
&\quad \times P(r|\lambda_r(r_d^b, r_{d0\ldots dU}^a, \theta_{d0\ldots dT})) \\
&\geq \mathbb{E}_{q(W)}\Big[\log P(W|\alpha_W, \beta_W) - \log q(W)\Big] \qquad \text{(B.1)} \\
&\quad + \mathbb{E}_{q(\mathrm{Q})}\Big[\log P(\mathrm{Q}|\alpha_c, \beta_c) - \log q(\mathrm{Q})\Big] \\
&\quad + \sum_{d \in D} \mathbb{E}_{q(\theta_d^p)}\Big[\log P(\theta_d^p|\alpha_c^p, \beta_c^p) - \log q(\theta_d^p)\Big] \\
&\quad + \sum_{d \in D} \mathbb{E}_{q(r_d^b)}\Big[\log P(r_d^b|r, 1) - \log q(r_d^b)\Big] \\
&\quad + \sum_{u \in U, d \in D} \mathbb{E}_{q(r_{du}^a)}\Big[\log P(r_{du}^a|r, 1) - \log q(r_{du}^a)\Big] \\
&\quad + \sum_{t \in d, d \in D} \mathbb{E}_{q(\theta_t^s)}\Big[\log P(\theta_t^s|\alpha_c^p, \beta_c^p) - \log q(\theta_t^s)\Big] \\
&\quad + \sum_{d \in D} \mathbb{E}_{q_1(\cdots)}\Big[\log P(r|\lambda_r(r_d^b, r_{d0\ldots dU}^a, \theta_{d0\ldots dT}))\Big] \qquad \text{(B.2)} \\
&\quad + \sum_{w \in t, t \in d, d \in D} \mathbb{E}_{q_2(\cdots) q(n_{wt})}\left[\log P\Big(n^{k^p k^s}(w,t)|W_{wk^p} \mathrm{Q}_{k^p k^s} \theta_{tk^s}^s \theta_{dk^p}^p\Big)\right] \\
&\quad + \sum_{w \in t, t \in d, d \in D} \mathbb{E}_{q_2(\cdots) q(n_{wt})}\left[\log P\Big(n^{k^s}(w,t)|W_{wk^s} \theta_{tk^s}^s\Big)\right]
\end{aligned}
$$

where $q_1(\cdots)$ is defined as,

$$
q_1(\cdots) = q(r_d^b) \prod_u q(r_{du}^a) \prod_k q(\theta_{dk}^p) \prod_{kt} q(\theta_{tk}^s)
$$

and $q_2(\cdots)$ is defined as,

$$q_2(\cdots) = q(W)q(Q)\prod_k q(\theta^p_{dk})\prod_{kt} q(\theta^s_{tk})$$

We can not derive the closed form update equations for Eq. B.1 due to the term Eq. B.2. At the same time, the term Eq. B.2 does not exist an analytic form of the expectation, but we can estimate the gradient of the term using the Monte Carlo method and applying reparameterization method (Naesseth et al., 2017) to reduce variance.

Appendix C

Topics from the Attribute Model

Table C.1 lists the top 20 words of the inter-sentence topics identified by the attribute model. Table C.2 lists top 20 words of the intra-sentence topics. We name the topics in the column titles by interpreting the top words. We notice that the inter-sentence topics often represent food types, while the intra-sentence topics often represent product attributes shared by the food types.

	Mexican	Breakfast	Italian	Steakhouse	Japanese	Dessert
0	taco	breakfast	bread	steak	roll	dessert
1	chip	egg	cheese	side	sushi	dress
2	mexican	pancake	sauce	cook	fish	chocolate
3	salsa	toast	pasta	meat	fresh	perfect
4	burrito	bacon	tomato	potato	chef	enjoy
5	margarita	brunch	italian	well	tuna	cheesecake
6	tortilla	french	dish	cut	salmon	cream
7	guacamole	morning	dip	medium	japanese	fan
8	enchilada	potato	garlic	salad	tempura	yum
9	cheese	hash	taste	rare	piece	cake
10	bean	waffle	oil	filet	spicy	treat
11	corn	coffee	butter	beef	sashimus	creamy
12	nacho	delicious	olive	tender	quality	yummy
13	asada	biscuit	serve	steakhouse	special	ice

Continued on next page

	Mexican	Breakfast	Italian	Steakhouse	Japanese	Dessert
14	fish	brown	meat	mash	miso	fabulous
15	carne	sausage	meatball	lobster	rice	sweet
16	chile	omelet	ravioli	appetizer	sake	apple
17	tamale	gravy	mushroom	mignon	soy	cookie
18	fajita	bagel	calamarus	juicy	ayce	vanilla
19	guac	syrup	marinara	seasoned	nigirus	strawberry

Table C.1: Inter-sentence topics

	Wait time	Service	Dining environment	Value	Family friendly	Taste
0	wait	table	room	price	friendly	taste
1	table	take	look	portion	husband	better
2	long	ask	wall	quality	dinner	dry
3	seat	server	atmosphere	small	reservation	ok
4	minute	waiter	light	size	family	bland
5	line	drink	dining	high	party	flavor
6	hour	arrive	kitchen	average	birthday	lack
7	sit	bring	enjoy	worth	server	nothing
8	busy	meal	music	nice	kid	expect
9	people	waitress	ambiance	dry	cup	much
10	around	check	next	better	hungry	salty
11	take	water	area	expensive	daughter	bad
12	want	first	find	meal	home	disappointed
13	area	wait	forward	little	accommodate	salt
14	min	seat	decor	think	water	bit
15	seating	friendly	stay	huge	need	sauce
16	worth	experience	beautiful	overall	right	cold
17	slow	sit	old	pay	attentive	overcooked
18	quickly	greet	smell	large	celebrate	mediocre
19	immediately	staff	inside	give	guest	tasteless

Table C.2: Intra-sentence topics

Appendix D

Website Design

We provide the screen-shots of the website used in the experiment. Participants were assigned to one of the sorting design groups after providing the preference. The participants were not aware of which group they were assigned to. Each participant reviewed three products and completed a survey after each product. After finishing the task, the website provided the participant with a unique code that allowed the participant to claim the reward in Amazon Mechanical Turk.

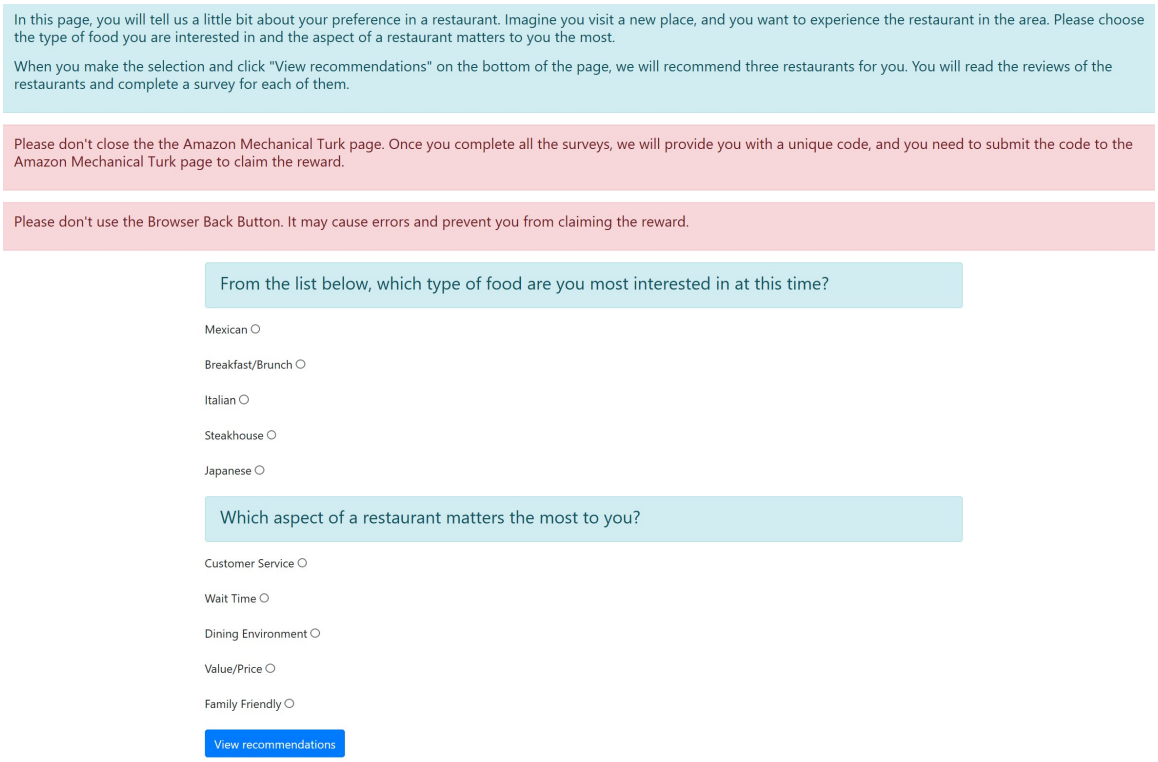

Figure D.1: Preference elicitation page

Figure D.2: Product presentation page 1

2015-01-02

One of the best tacos I have tasted. I'm Latina and I have ate a lot of tacos in my life but these ones for some reason remind me of home lol. For some reason the ones in Brookline taste better than the ones in beechview or Oakland. I also love that you are able to go to the store and buy Mexican sodas and drink them while eating your tacos. A true 3rd world country experience in the burgh!

Stars: 4

2014-12-30

This place is seriously great. It's a little stand next to the grocery store. I learned about it from my roommate who is from Mexico. She said it was the real deal, so I had to check it out. Delicious tacos for two bucks! The beef are my favorite. They also have a wide assortment of sauces that are all a little too spicy for me. I get mine plain, but they are just phenomenal.

Stars: 4

2014-11-25

While nearby and looking for a quick meal, I stopped by the outdoor taco stand to pick up a few tacos to go. Despite the chilly weather, there was a line for the tacos. As I waited, I watched the tacos being made. The meats are precooked and kept warm. When you order, the corn soft shells are heated and the meat of your choice is added. When I asked the options, I struggled to understand what was communicated to me so I simply asked for 2 pork, 2 ground beef and 2 chicken. However, there were about 6 options including what I believed to be steak and a fajita mix. I then went inside to the toppings bar and added what I believed to be Pico de Gallo and a creamy guacamole, unfortunately nothing was labeled. I tried to ask a few questions to the cashier but she had some difficulty understanding me. Seems they could really benefit from a few posted signs with the meat options, prices and toppings. I took the tacos to go and shared them with my family. Overall, I felt the taste was authentic though the meat was a bit fatty for me. I also felt it lacked some overall flavor. For $2.50 a taco, it was a great deal but I would not go out of my way to return.

Stars: 3

2014-11-03

OmiGOD their tacos are freaking insane. They're better than those I'd had when I was in California, seriously. Great. Tacos. Also they have pretty cheap produce inside and it's not bad! I generally wait to get my onions so that I can pick them up here. Potatoes are cheap too. lololol one time I went there for graham crackers and I was trying to explain to the lovely and friendly cashier who does NOT speak English what I wanted. She called another guy over and I tried to explain to him what I was looking for. Then he told me that if I made "sweet chicken" it wouldn't taste good. LOL. I found cinnamon cookies and ran out of there, slightly embarrassed, slightly amused.

Stars: 5

2014-10-12

Tacos all day, errday. Tacos from breakfast, tacos for lunch, tacos for dinner, tacos for midnight snack. I could eat every single one of their tacos (and I have) and I can because it's only 2.50 a taco. I've only set foot in their grocery store to pay for my tacos.

Stars: 4

Previous page of reviews 1 of 12 Next page of reviews

Continue to the survey

Figure D.3: Product presentation page 2

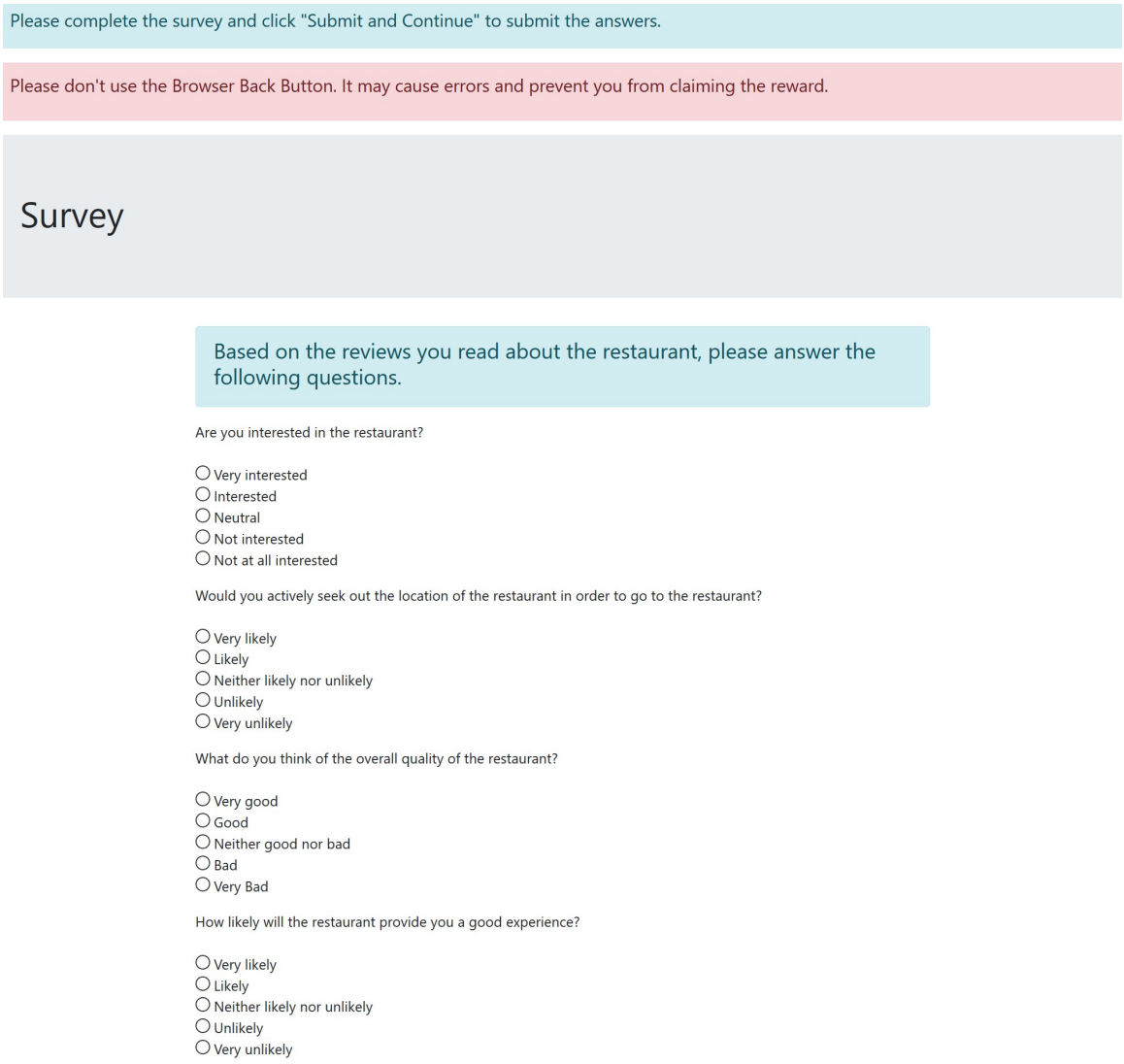

Figure D.4: Survey page 1

◯ Very likely
◯ Likely
◯ Neither likely nor unlikely
◯ Unlikely
◯ Very unlikely

Would you like to try the restaurant?

◯ Very likely
◯ Likely
◯ Neither likely nor unlikely
◯ Unlikely
◯ Very unlikely

Would you like to visit the restaurant when you happen to see it?

◯ Very likely
◯ Likely
◯ Neither likely nor unlikely
◯ Unlikely
◯ Very unlikely

Would you actively seek out the location of the restaurant in order to go to the restaurant?

◯ Very likely
◯ Likely
◯ Neither likely nor unlikely
◯ Unlikely
◯ Very unlikely

Now please consider your answers about the restaurant, are you agree with the following statements?

My answers were based on the best available information.

◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

My answers were made based on valid assumptions.

◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

Figure D.5: Survey page 2

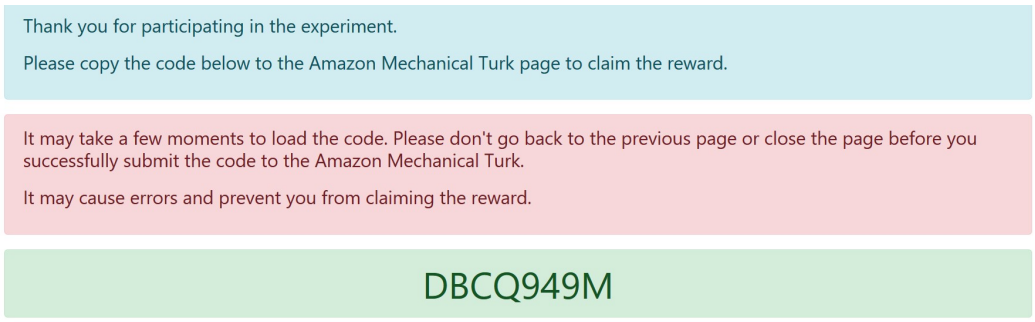

Figure D.6: Reward claim page

Appendix E

Top Five Reviews of the Sorting Designs

We provide an example of the top five reviews of three sorting designs for the recommended product and the bad recommendation. The food type of interest is "Mexican food," and the restaurant aspect of interest is "Price/Value." The personalization computes the recommended product as "Las Palmas Carniceria" and the bad recommendation as "Phat Phrank's". Both restaurants have a 4.5 overall rating. Table E.1 lists the top five reviews of time sorting for the recommended product and the bad recommendation. Table E.2 lists the top five reviews of the recommended product in vote sorting and personalized sorting. Three out of the top five reviews of the recommended product overlap between the two sorting designs, which indicates, for the recommended product, the helpfulness votes align with the customer's preference. The top reviews from both sorting designs discuss the exceptional value the restaurant offers. However, for the bad recommendation, the top five reviews of two sorting

designs have no overlap (see Table E.3). The reviews from the personalized sorting design focus on the discussion that the restaurant is overpriced, while the reviews from the vote sorting design emphasize the high-quality of the food and service. The example is consistent with our analysis of the vote sorting design that it is helpful for the customer when the product matches the customer's preference but less effective when the product does not match the customer's preference.

No.	Recommended product	Bad Recommendation
1	One of the best tacos I have tasted. I'm Latina and I have ate a lot of tacos in my life but these ones for some reason remind me of home lol. For some reason the ones in Brookline taste better than the ones in beechview or Oakland. I also love that you are able to go to the store and buy Mexican sodas and drink them while eating your tacos. A true 3rd world country experience in the burgh!	Came here because of a classmates recommendation. Ordered the barbacoa burrito. Pros-speedy service, flavorful burrito. I recommend the mild sauce. A great mom&pop store (5 stars). Will go to again. Cons- the spicy sauce is very, very spicy. I wish there was a scale from 1 to 10 for more options instead of only 1 option. However, I was impressed with their level of spice. (4.5 stars)
2	This place is seriously great. It's a little stand next to the grocery store. I learned about it from my roommate who is from Mexico. She said it was the real deal, so I had to check it out. Delicious tacos for two bucks! The beef are my favorite. They also have a wide assortment of sauces that are all a little too spicy for me. I get mine plain, but they are just phenomenal.	Really cute family place with great prices. We loved the food and took their recommendations on what to order. Mr R got a breakfast burrito with carne asada and I got the adobada plate. Loved them both. The pork especially was very good with the green salsa.

No.	Recommended product	Bad Recommendation
3	While nearby and looking for a quick meal, I stopped by the outdoor taco stand to pick up a few tacos to go. Despite the chilly weather, there was a line for the tacos. As I waited, I watched the tacos being made. The meats are precooked and kept warm. When you order, the corn soft shells are heated and the meat of your choice is added. When I asked the options, I struggled to understand what was communicated to me so I simply asked for 2 pork, 2 ground beef and 2 chicken. However, there were about 6 options including what I believed to be steak and a fajita mix. I then went inside to the toppings bar and added what I believed to be Pico de Gallo and a creamy guacamole, unfortunately nothing was labeled. I tried to ask a few questions to the cashier but she had some difficulty understanding me. Seems they could really benefit from a few posted signs with the meat options, prices and toppings. I took the tacos to go and shared them with my family. Overall, I felt the taste was authentic though the meat was a bit fatty for me. I also felt it lacked some overall flavor. For $2.50 a taco, it was a great deal but I would not go out of my way to return.	I have been eating hear since they opened and it's been consistently excellent to this day. My favorite is the Chile rellano burrito. But I've tried about everything and it's all good. Family run and friendly place.
4	OmiGOD their tacos are freaking insane. They're better than those I'd had when I was in California, seriously. Great. Tacos. Also they have pretty cheap produce inside and it's not bad! I generally wait to get my onions so that I can pick them up here. Potatoes are cheap too. lololol one time I went there for graham crackers and I was trying to explain to the lovely and friendly cashier who does NOT speak English what I wanted. She called another guy over and I tried to explain to him what I was looking for. Then he told me that if I made "sweet chicken" it wouldn't taste good. LOL. I found cinnamon cookies and ran out of there, slightly embarrassed, slightly amused.	Came from LA to Vegas and we wanted Mexican food. Yelped it and came here. Great tortas, burritos and taquitos. The beans didn't taste homemade.
5	Tacos all day, errday. Tacos from breakfast, tacos for lunch, tacos for dinner, tacos for midnight snack. I could eat every single one of their tacos (and I have) and I can because it's only 2.50 a taco. I've only set foot in their grocery store to pay for my tacos.	Awesome place for lunch in the area. These people know good Mexican cuisine. I had the carnitas burrito with red sauce. Chili sauce "wet" is the way to go, trust me. Carnitas were tender and juicy. Also, it's family owned/operated. Nothing "commercial" about this place at all. Just real, flavorful food.

Table E.1: Top five reviews of a pair of recommended product and bad recommendation in time sorting

No.	Vote sorting	Personalized sorting
1	I had heard rumors about these delicious tacos. But I didn't really fully understand it until I had them today. Today I fully understood their deliciousness. It's such a simple thing. One stands at taco stand. One looks over delicious looking meat. One chooses between a couple different delicious looking meat including steak and onions, sausage and pork. One receives taco and walks over to a neat and clean looking buffet of delicious looking toppings. These toppings include salsa, salsa verde, gauc, tomatoes and onions. One devours tacos in an animal like fashion because... they're delicious! So for serious, they rocked my world! And at $2 a pop you can buy a ton without breaking the bank! I had the steak and the pork. I liked the pork more because it was uber flavorful. I also loved the toppings bar because I LOVE to dump toppings on as well. Although at two tacos I stuffed myself and was full for hours (which does not usually happen!). Do they beat out Reyna's tacos? My answer would be they're on par but on different levels because they're just a different style. That's just my observation. The more taco stands, the merrier if you ask me. Bring it tacos!	Las Palmas is the platonic ideal of the ultimate mexican taco stand. I'm normally not even a taco fan, but they take it to a completely different level, every savory bite is a transformative flavor experience. Here is Las Palmas by the numbers: 1 other mexican restaurant's all-mexican staff that told me Las Palmas is where THEY all go out to eat 2 dollars is the price of each taco. Each taco is actually wrapped with 2 hand made soft taco tortillas and have a double portion of meat (potato salad is an option for vegetarians), so it's more like 1 dollar per standard size taco. Gourmet, fresh, delicious mexican at cheaper than taco bell prices! 4 is how many tacos I ate. I ordered 2, left, ate them, and immediately returned to buy more to go to bring home for dinner that night! 10 is how many different fresh toppings they have, guacamole, peppers, limes, cilantro, onions and a variety of salsas from mild to absolutely DEADLY hot habaneros! 7 is how many days a week I'd happily eat Las Palmas tacos!

No.	Vote sorting	Personalized sorting
2	Finally, after four years in Pittsburgh, I found a real Hispanic grocery store, and it comes with a real taco stand. I've been grudgingly going to a well-known place in the Strip that charges 20-50% more for a not-so-great selection, but I shall shop there no more now that I've found Las Palmas. Fully stocked grocery store with great produce (huge pile of tomatillos) and meat - everything you need to make that recipe you found in your favorite Diana Kennedy cookbook. (Cheap Inca Kola! Found my fix.) Most of you are more interested in the tacos, and I'll confirm that you can't go wrong there, either. Real tacos - I tried puerco, barbacoa, and asada. All were very good; the pork was exceptional - muy rica. These are the real deal for street tacos, too - just like what I found in Atlanta and San Diego in past lives. Only $2 (never pay $3 for a taco), and they're stuffed with easily twice as much fillin's as what you'll get at any other taco place in town. (Note to their competition: tacos shouldn't be huge, but they're not appetizers, either. I'm tired of eating three at the trendier places and still being hungry. Four with no sides should completely stuff an adult male.) The "fixin's" station is nice, too - faster and easier than having to wait on "the guy" to put it on. Four stars for the grocery store and five for the taco stand. (...And the grocery store takes credit cards. Hint, hint to all the local shops/restaurants that don't.)	Finally, after four years in Pittsburgh, I found a real Hispanic grocery store, and it comes with a real taco stand. I've been grudgingly going to a well-known place in the Strip that charges 20-50% more for a not-so-great selection, but I shall shop there no more now that I've found Las Palmas. Fully stocked grocery store with great produce (huge pile of tomatillos) and meat - everything you need to make that recipe you found in your favorite Diana Kennedy cookbook. (Cheap Inca Kola! Found my fix.) Most of you are more interested in the tacos, and I'll confirm that you can't go wrong there, either. Real tacos - I tried puerco, barbacoa, and asada. All were very good; the pork was exceptional - muy rica. These are the real deal for street tacos, too - just like what I found in Atlanta and San Diego in past lives. Only $2 (never pay $3 for a taco), and they're stuffed with easily twice as much fillin's as what you'll get at any other taco place in town. (Note to their competition: tacos shouldn't be huge, but they're not appetizers, either. I'm tired of eating three at the trendier places and still being hungry. Four with no sides should completely stuff an adult male.) The "fixin's" station is nice, too - faster and easier than having to wait on "the guy" to put it on. Four stars for the grocery store and five for the taco stand. (...And the grocery store takes credit cards. Hint, hint to all the local shops/restaurants that don't.)

No.	Vote sorting	Personalized sorting
3	Las Palmas is the platonic ideal of the ultimate mexican taco stand. I'm normally not even a taco fan, but they take it to a completely different level, every savory bite is a transformative flavor experience. Here is Las Palmas by the numbers: 1 other mexican restaurant's all-mexican staff that told me Las Palmas is where THEY all go out to eat 2 dollars is the price of each taco. Each taco is actually wrapped with 2 hand made soft taco tortillas and have a double portion of meat (potato salad is an option for vegetarians), so it's more like 1 dollar per standard size taco. Gourmet, fresh, delicious mexican at cheaper than taco bell prices! 4 is how many tacos I ate. I ordered 2, left, ate them, and immediately returned to buy more to go to bring home for dinner that night! 10 is how many different fresh toppings they have, guacamole, peppers, limes, cilantro, onions and a variety of salsas from mild to absolutely DEADLY hot habaneros! 7 is how many days a week I'd happily eat Las Palmas tacos!	I stopped by here this past Tuesday afternoon for lunch with my father. I had noticed this place the week before whenever we went by and became eager to try it whenever I noticed its high rating on yelp. Since I didn't go into the store, the only real thing I can comment on is the taco stand outside. There was a guy behind the stand cooking everything as well as a condiment bar next to it. Whenever we walked up he was just finishing up grilling some tortillas. He greeted us in Spanish before giving us a run down of the taco fillers in English. I ended getting three, one each of steak, chorizo and chicken. He quickly made the tacos stuffing them very full. They consisted of two tortillas with the filling which reminded me of truly authentic tacos. Whenever I asked about paying he informed that I could pay cash outside or credit inside. Since I had cash I quickly payed him and moved over to the condiment bar. They had a number of sauces as well as some onions and a few other veggies. I ultimately ended up choosing what I think was an avocado based sauce. Once we were done with the condiments my father and I walked down the boulevard to find a public bench to eat our tacos on. After sitting down I quickly dug in. All three of them were delicious! My favorite of the bunch ended up being the steak, followed by the chicken and lastly the chorizo. The chorizo was the lone disappointment of the meal as I feel it lacked any spice at all but it was still pretty tasty. The bottom line for me is that I will absolutely head back here again in the future. The tacos were all really good and I think that they were a great value at $2.50 each.

No.	Vote sorting	Personalized sorting
4	Thank You Las Palmas. Thank you for answering the unsatisfied craving for quality authentic West Coast Mexican food I've had for years since leaving Orange County. Thank you for being that cheap (read $6 for 3 tacos) staple quick food fix when I don't feel like taking the car or bike out to grab some grub. And thank you for offering not only a great taco stand with a variety of fresh simple dressings, but also an always fully and freshly stocked ethnic grocery store where I can find hard to get items for a quarter of the price that competitors charge. As I've watched you expand over the past 3 years... just PLEASE do me a favor, don't lose sight of what makes you great!	Short Version: Great selection of Mexican food, as well as fresh produce and other grocery items, at prices that are more than reasonable. Long Version: Having moved here recently from Southern California, seeing a Carniceria open down the street was an exciting prospect, but I expected to pay for that privilege. I was wrong: prices are as cheap, or in some cases (fresh bell peppers, some produce and butcher items) cheaper, as you'd find at Costco or Giant Eagle. Plus, they have all the goodies you want like Coke bottled in Mexico (real sugar! no corn syrup), Jarritos Mexican soda, and a large assortment of those delicious and hard to find items you might associate with Mexican food or things available in Southern California. This includes pre-seasoned burrito and taco meats like Al Pastor for very cheap prices. The only imperfect thing about the whole operation is that some items that are stocked are stocked in small numbers, so occasionally you'll go back to find they're out of stock of something you were looking for, but it'll usually be there next time you go back. Highly Recommended
5	Lissen up folks! You're really not going to get much better than this Las Palmas location. The prices are extremely good, the taco stand is awesome, and they have a full deli. I am not aware of any other full service grocerias in Pittsburgh. The selection is quite complete - you can get anything from a selection of dried chilis to Salvadorian crema, and the produce is well priced as well! There's even a full cooler if you want to buy some agua con gas or Jarritos! When I need to find real ingredients (they also regularly carry fresh cactus paddles..) this is the true stop in da burgh.	Lissen up folks! You're really not going to get much better than this Las Palmas location. The prices are extremely good, the taco stand is awesome, and they have a full deli. I am not aware of any other full service grocerias in Pittsburgh. The selection is quite complete - you can get anything from a selection of dried chilis to Salvadorian crema, and the produce is well priced as well! There's even a full cooler if you want to buy some agua con gas or Jarritos! When I need to find real ingredients (they also regularly carry fresh cactus paddles..) this is the true stop in da burgh.

Table E.2: Top five reviews of a recommended product

No.	Vote sorting	Personalized sorting
1	Phat Phrank's has been on my 'To Do List' for a very long time. I had a meeting on the other side of town and had to pass by the Decatur & 215 exit on the way home around lunchtime, so I decided to finally give Phat Phrank's a try. Since I'm on my Lent 'No Meat' diet, I was browsing the menu for Meatless items. The man behind the register asked me if he could help out....I told him my situation and he was very happy to help me decide what to get. He suggested a Cheese Enchilada ($6.99). I went with his suggestion and also added a Bean & Cheese Burrito ($4.29) for good measure.....ha! He chatted me up while ringing up my order and I discovered he was Phrank, the owner. He's such a cool guy. He's from So Cal and we were talking a bit about the State as I'm also a So Cal transplant.....although I left many years before he did. After a few minutes he brought out my order. I was pleased to see that the Cheese Enchilada also came with a side of Rice & Beans....although it meant that I ordered too much food. Sometimes my stomach tricks me into these situations. Oh well. I went for the Cheese Enchilada first....and yum! It was very, very good. I'm glad Phrank suggested it to me. Next up was the Burrito.....unfortunately, after all of the flavor from the Enchilada, the Burrito tasted just OK. For a Bean Burrito though, it was solid. Next time I'm going to order just one or the other. I'll have to come back after Lent so I can try out some of the other signature dishes. If they're as good as the Cheese Enchilada, I'll be back a lot. Return Factor - 97%	This was an upscale version of a taco shop. For the most part same taco shop fare but higher prices and more adjectives in their menu, such as, "flawless" and "delectable". The owner took our order. I asked him what he recommended and he said people like the fish taco and the tortas. I then asked him what his favorite was and he replied "I like them all it is like picking a favorite granchild" to which I said "but everyone has a favorite." He eventually told me until the article last week (which I have no idea of) barbacoa was the most popular. So I tried the two taco plate, one fish and one barbacoa. The barbacoa was a little dry and not a lot of flavor, also the portion of barbacoa for the price in the taco was pretty small. The fish taco was pretty good, fried cod I think, it was a pretty big piece of fish and it was fried well, a little greasy, but crispy. The plate also came with a small portion of rice and beans. I got the hot salsa - owner said not as hot as habaneros, but it still had a pretty good kick. Also, a medium Jamaica drink (free refills) - I paid a little under $12 - which I think is expensive for a taco shop. The food came on paper plates, which I would expect, but again for the price it seems like you should get more. It just didn't seem that authentic - but I am more of a Tacos El Gordo type of girl personally.

No.	Vote sorting	Personalized sorting
2	Phat Phrank's has quickly become my go-to place to tasty Mexican here in Vegas. It's a family owned place where the food is made with a lot of care and I love that! "Phrank" is always around and smiling and his staff will do all that they can to ensure your satisfaction. At this point I've tried quite a few things... my favorites: *Nachos - I'm actually not a nachos person, but his are done the way nachos should be! Tons of delicious shredded pork... an order is definitely a meal for 2 people! *Breakfast Burrito - the best breakfast burrito I've had in Vegas, by far! It comes with eggs, yummy crispy potatoes, cheese, and your choice of meat. He mixes everything to ensure every bite is perfect. When I'm having a bad day, I call ahead and pick one up on my way to work! *Fajitas - I have FINALLY found delicious fajitas in Vegas! I had the chicken ones for lunch today and really enjoyed them. The chicken was shredded (as opposed to the creepy pre-cut strips with faux-grill marks you find at say Chili's) and had 3 colors of peppers and onions mixed in. I'm not usually a fan of chicken, but this had an amazing amount of flavor. *Churros - caramel and chocolate sauce.. mmmmmmm I can also highly recommend their catering services ! "Phrank" helped us pick out a freaking feast for well under $10 a person. We had platters of enchiladas and tacos, guacamole, and ceviche. Everything was delicious and was ready right on time. We picked it up ourselves, but it was well packaged and transport -ready. Overall, don't let the strip mall exterior fool you - this is Mexican (albeit heavily Americanized) at its best ! It's affordable, fast, and delicious. It's basically like Roberto's that isn't a chain that serves food that tastes good! The only sad thing is they aren't open on weekends...	Had Phat Franks for the first time just a few hours ago. Very impressed with the food and Frank. He and his helper (didn't catch her name) are very cordial and really make an effort to offer exceptional service. Ordered one carnitas and two chicken tacos. They were phenomenal! I love tacos and try to sample as many as possible. Frank does a great job. The meat is flavorful, moist and plentiful, the corn tortillas fresh and the toppings are near salad like proportions (cabbage, cilantro and onion). I opted for the spicy salsa (there is a mild, green and spicy) which had a nice kick to it. There is no salsa bar, your choice is added to the order. The price, as I recall, is about $2.89 per taco. A little on the higher side, but they are worth it as the overall size and meat portion trumps the other tacos I've come across. The place is also very clean and the food is made to order and prepared quickly. Phat Frank's is simply PHantastic!

No.	Vote sorting	Personalized sorting
3	There's only one word I can think of when it comes to Phat Phrank's. AMAZING. I could leave it as that and have you all wonder or I can let you all know why it's so AMAZING. I've had their fish taco plate, other than the beans and rice sides the fish tacos were AMAZING. By far the best fish tacos I've had anywhere. The fish were large crispy on the outside tender on the inside pieces, light and not greasy. The combination of all the fish with cabbage, green onion, cilantro, tartar sauce, and in my case lime and their hot sauce was AMAZING. My mouth is watering as I reminisce. I'll be back for you fish tacos. Their carne asada torta was... Oops sorry for the drool. Their Carne asada torta was very delicious too. Well I might have reached my quota for using the word amazing but don't worry the carne asada tortas are equally as amazing as the fish tacos. I've seen other tortas that are bulky, bready, and just look good because of the huge size not flavor. Size isn't gigantic and not bready, so every bite was just of everything I wanted, something very tasty. There's probably even more ingredients than bread. I'll be back for you too LOL. I didn't have this but a friend had their enchiladas because it was mentioned in the lvrj as Phat Phrank's signature dish. Now that looked like meal right there. I'm sure it has to be really good too just from what I've already had. Plus it had a fried egg on it. Anytime you throw a fried egg on anything it's instantly good. Location is perfect for me, I'm always in the area though honestly I never knew they were there until a UYE. I like the musical instrument on the walls and how it's clean. Seats maybe 20-30 people. Service was very good, friendly. Phrank and the lady that was there often checked on us, refilling drinks, bring food out quickly, conversing with us and making it an experience not like here is your food and now get out when your done. lol. They no longer have raspberry ice tea. Can't wait to go there tomorrow. Wait their not open on the weekends!? This has to be a typo and they close at 6 PM. MAN I can't win. All good, fresh and delicious food need a break too. AMAZING.	I had the adobada torta with the and the fish taco. Torta was amazing, however I think it could have used a bit more meat. The fried cod taco I think cost too much, however the fish and sauce are delicious, just not worth the price and my corn tortilla fell apart. Good service in a shopping center restaurant. maybe come back to try the desserts.

No.	Vote sorting	Personalized sorting
4	Oh Phrank, I've got the hots for you! My hubby tried Phat Phrank's last week with some coworkers and has been raving about the fish tacos every day since then. I wanted in on the action, so when he had the opportunity to go out for lunch again, I met him there and had a great lunch. The restaurant itself is in a small strip mall with a couple of other food places. It is kind of hard to see from the street, so keep that in mind when you venture over here. Once you find Phrank's, your happy tummy time begins! You walk in the door and can smell the lovely Tex-Mex aromas swirling around the place. Your mouth starts watering immediately and you get excited for the food to come. Yummmm. Phat Phrank's is a small place though, with about 6 tables. We got there and there was only 1 table open. My hubby grabbed it while I waited in the really long line. Within 2 minutes, the line was actually out of the door with people waiting. Luckily the line moves pretty fast. The food also comes out pretty quickly too, so there is a nice flow to the place. After looking at the menu, I was torn between the carne asada and the pork chili verde plates. I asked the one and only Phrank, who was working the counter, and he told me to go for the chili verde. I trusted him and went with the verde. Boy, was I happy with my choice! The pork chili verde was so tender and tasty. It had so many layers of flavor that it made you eat slowly to savor each bite. Delish! My hubby went for the fish tacos again. He is addicted to fish tacos and gets them everywhere we go. He says these are the best in town, hands down. He really has tried dozens and dozens, so I trust him on fish tacos. I was going to try a bite, but by the time I asked for one, he had already devoured them. You snooze you lose, I guess. The hot sauce here is really great good too. I recommend the hot hot sauce. It is super flavorful without being completely mouth burning. It has some great depth to it and I had to ask for a second serving because it was that good! The plates came with the main entree, as well as refried beans, Mexican rice, and homemade tortillas. Now no one tell my Grandma, but the beans and tortillas were better than hers! I have NEVER said that before because hers are the best of the best! The beans from Phrank's were simply perfect and I could have eaten a whole pot of them them. The flour tortillas tasted just like my grandma's homemade ones. The tortillas were just AWESOME. If you have never had a homemade tortilla, you are missing out. Luckily you can go to Phrank's and get some right now!	I'm not sure why this place has so many great reviews. The food was edible. That's about it. Nothing stood out. Personally, I think I could get a more authentic fish taco at Rubio's and a better tasting carnitas burrito at Chipotle–and at 1/2 the price. There wasn't enough flavor or spice or anything, really. Was asked what kind of salsa do I want? Um, I don't know, how about all of them? How do I know which one is best if I can't try them all? Total turn off. My burrito was dry and I didn't like having to beg for more soda. For God's sake invest in a salsa bar & let people get their own soda! One more tip: don't run out of utensils. Especially when customers don't appear to be lining up. When business is that slow, you shouldn't really run out of anything or make your guests wait while you restock the forks.

No.	Vote sorting	Personalized sorting
5	FIVE stars. Why? Two words: FISH TACOS. Also, three more words: PERSONABLE, FRIENDLY OWNER. I didn't think this place was open after 6pm, which is part of the reason why I haven't tried their food yet. However, we left Halloween Mart, drove by, and the lights were on and the OPEN sign was lit up! SCORE. Apparently, on THURSDAYS they are open until 9pm because they recently started having some type of car meet up in the parking lot. NOW. THE FOOD. The inside is nothing special - not much atmosphere or ambiance. We took a look at the menu and the man behind the counter, who I assume is the owner, came over to us with a huge smile and said, "HOW CAN I HELP YOU LADIES?" We told him we were first timers. He explained how a few of the menu items are cooked and told us what his favorites were. As soon as he said FISH TACOS, my eyes lit up. I ordered one fish taco and one adobada taco with rice and beans. My girlfriend ordered one carne asada taco and one adobada taco - TO GO. While we were waiting, the same nice man came and chatted with us for a while, asking us where we were from and he actually engaged in some awesome and genuinely friendly conversation. I was very impressed with his hospitality! Our food was ready in a few minutes and as soon as I got home, I took one bite into that fried, crispy, juicy, flaky, white fish topped with lettuce and cilantro and a creamy sauce, and I just about died right then and there. Died with pleasure of course. :) Honestly, probably the second best fish taco I have ever had in my entire life, and I don't play around when it comes to fish tacos. The adobada meat was slightly dry and lacked a bit in flavor. The beans and rice were amazing. i can't quite put my finger on the spice/flavor that made them taste different, but there's definitely something special about them. Regardless, I cannot give less than 5 stars because of the beautiful explosion of taste that came out of that fish taco. I will be back!	Usually I run a mile when faced with inexpensive Mexican food, but the overall rating on Yelp for Phat Phrank's convinced me to give it a try. I'm glad I did. This is good, honest Mexican fare at a really respectable price. It won't burn a hole in your wallet and it's also not as unpredictable as the stuff served up and places like Roberto's, Don Tortaco, Cafe Rio, and others. Everything has the feel of being freshly prepared rather than slopped together out of vats and it tastes just as good. I really appreciated the texture of the tortillas which were warm and satisfying, dense and chewy enough so they don't split open in your hands, but not gluey and overwhelming, either. The man himself, Frank, is among the most welcoming proprietors I have encountered anywhere in the Las Vegas valley. He took a genuine interest in me and my wife and remembered my name upon my next visit - admittedly this was the very next day, but that should stand as testimony to the impact of his hospitality and cooking. This isn't one of those fancy, designer Mexican restaurants, and should not be held up against that sort of establishment - these two sub-categories of Mexican cuisine can barely even be considered the same type of food. For what he is aiming for, Frank scores a bullseye. He deserves his high Yelp rating.

Table E.3: Top five reviews of a bad recommendation

Appendix F

Correlation, Normality and Homoscedasticity in Measurements

We examine the correlation among the questions measures the same constructs to ensure the validity. The results in Table F.1 and Table F.2 show a good correlation among these questions.

We use the D'Agostino-Pearson test to examine whether to reject the hypothesis that the data of a variable is normal. The original data of the time used, the number of words read, and their related variables are rejected by the test. The plots of the total time used and the total number of words read suggest that the variables may have a log-normal distribution. We transform the original data to the logarithm that the test can no longer reject the null hypothesis. We list the p-values of the test in Table F.3. To satisfy the assumption of ANOVA, we use Bartlett's test to examine the equality of variances. The test can not reject the null hypothesis that their variances are equal. The p-values of the test are in Table F.4.

	Q0	Q1	Q2	Q3	Q4	Q5	Q6
Q0	1.00	0.83	0.82	0.76	0.76	0.75	0.72
Q1		1.00	0.84	0.79	0.80	0.78	0.75
Q2			1.00	0.81	0.80	0.79	0.77
Q3				1.00	0.89	0.87	0.84
Q4					1.00	0.91	0.85
Q5						1.00	0.84
Q6							1.00

Table F.1: Measurements of customer's intention: Q0-Q2 measure the perceived quality and Q3-Q6 measure the purchase intention.

	Q7	Q8	Q9	Q10	Q11	Q12
Q7	1.00	0.75	0.77	0.70	0.71	0.73
Q8		1.00	0.79	0.72	0.70	0.71
Q9			1.00	0.7	0.71	0.73
Q10				1.00	0.73	0.75
Q11					1.00	0.79
Q12						1.00

Table F.2: Measurements of decision quality: Q7-Q9 measure the perceived decision quality and Q10-Q12 measure the decision confidence.

Variable Name	Test for Normality (p-value) Time	Sort	Personalized
Intention difference		0.77	
Time used difference (in log)		0.17	
# of words read difference (in log)		0.20	
Intention for recommended products	0.61	0.94	0.16
Intention for bad recommendations	0.39	0.36	0.67
Intention for random products	0.96	0.65	0.58
Intention difference	0.95	0.19	0.79
Decision quality for recommended products	0.19	0.16	0.17
Decision quality for bad recommendations	0.36	0.19	0.12
Decision quality for random products	0.15	0.19	0.11
Total time used (in log)	0.45	0.13	0.14
Time used in examining recommended products (in log)	0.72	0.82	0.11
Time used in examining bad recommendations (in log)	0.34	0.27	0.21
Time used in examining random products (in log)	0.96	0.38	0.82
Total # of words read (in log)	0.17	0.26	0.36
# of words read of recommended products (in log)	0.14	0.20	0.41
# of words read of bad recommendations (in log)	0.34	0.95	0.51
# of words read of random products (in log)	0.15	0.19	0.61

Table F.3: Normality test

Variable Name	Test for Equality of Variances (p-value)
Intention for recommended products	0.57
Intention for bad recommendations	0.26
Intention for random products	0.73
Intention difference	0.16
Decision quality for recommended products	0.83
Decision quality for bad recommendations	0.61
Decision quality for random products	0.69
Total time used (in log)	0.24
Time used in examining recommended products (in log)	0.34
Time used in examining bad recommendations (in log)	0.16
Time used in examining random products (in log)	0.61
Total # of words read (in log)	0.14
# of words read of recommended products (in log)	0.19
# of words read of bad recommendations (in log)	0.13
# of words read of random products (in log)	0.17

Table F.4: Equality of variances test

References

Ansari, A., Essegaier, S., and Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research*, *37*(3), 363–375.

Aral, S., Dellarocas, C., and Godes, D. (2013). Introduction to the special issue—Social media and business transformation: A framework for research. *Information Systems Research*, *24*(1), 3–13.

Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485–1509.

Bagheri, A., Saraee, M., and De Jong, F. (2014). Adm-lda: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, *40*(5), 621–636.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baker, M. J., and Churchill Jr, G. A. (1977). The impact of physically attractive models on advertising evaluations. *Journal of Marketing Research*, 538–555.

Bell, R. M., Koren, Y., and Volinsky, C. (2010). All together now: A perspective on the Netflix Prize. *Chance*, *23*(1), 24–29.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Benlian, A. (2015). Web personalization cues and their differential effects on user assessments of website value. *Journal of Management Information Systems*, *32*(1), 225–260.

Benlian, A., Titah, R., and Hess, T. (2012). Differential effects of provider recommendations and consumer reviews in e-commerce transactions: An experimental study. *Journal of Management Information Systems*, *29*(1), 237–272.

Bernstein, F., Modaresi, S., and Sauré, D. (2018). A dynamic clustering approach to data-driven assortment personalization. *Management Science*, *65*(5), 2095–2115.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 242–262.

Bickart, B., and Schindler, R. M. (2001). Internet forums as influential sources of consumer information. *Journal of interactive marketing*, *15*(3), 31–40.

Blei, D. M., and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 17–35.

Blei, D. M., and Mcauliffe, J. D. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems* (pp. 121–128).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boureau, Y.-L., Ponce, J., and LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 111–118).

Brady, M. K., and Cronin Jr, J. J. (2001). Some new thoughts on conceptualizing perceived service quality: a hierarchical approach. *Journal of Marketing*, *65*(3), 34–49.

Branco, F., Sun, M., and Villas-Boas, J. M. (2012). Optimal search for product information. *Management Science*, *58*(11), 2037–2056.

Branco, F., Sun, M., and Villas-Boas, J. M. (2015). Too much information? information provision and search costs. *Marketing Science*.

Büschken, J., and Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953–975.

Cao, Q., Duan, W., and Gan, Q. (2011). Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. *Decision Support Systems*, *50*(2), 511–521.

Chen, P. Y., Dhanasobhon, S., and Smith, M. D. (2008). *All reviews are not created equal: The disaggregate impact of reviews and reviewers at Amazon.com.* Retrieved from https://ssrn.com/abstract=918083

Chen, Y., and Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, *54*(3), 477–491.

Chen, Y. C., Shang, R. A., and Kao, C. Y. (2009). The effects of information overload on consumers subjective state towards buying decision in the internet shopping environment. *Electronic Commerce Research and Applications*, *8*(1), 48–58.

Cheng, Z., Ding, Y., Zhu, L., and Kankanhalli, M. (2018). Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the 2018 World Wide Web Conference* (pp. 639–648).

Chevalier, J. A., and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Chintagunta, P. K., Gopinath, S., and Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944–957.

Clemons, E. K., Gao, G. G., and Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems*, *23*(2), 149–171.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological bulletin*, *104*(2), 163.

Cui, G., Lui, H. K., and Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce*, *17*(1), 39–58.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.

Davis, J. M., and Agrawal, D. (2018). Understanding the role of interpersonal identification in online review evaluation: An information processing perspective. *International Journal of Information Management*, *38*(1), 140–149.

Decker, R., and Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, *27*(4), 293–307.

Dellarocas, C., Zhang, X. M., and Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45.

Deloitte. (2014). *The Deloitte consumer review: The growing power of consumers.* Retrieved from https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/consumer-review-8-the-growing-power-of-consumers.pdf

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*(1), 31–71.

Dooley, R. S., and Fryxell, G. E. (1999). Attaining decision quality and commitment from dissent: The moderating effects of loyalty and competence in strategic decision-making teams. *Academy of Management Journal*, *42*(4), 389–402.

Fader, P. S., and Hardie, B. G. (1996). Modeling consumer choice among SKUs. *Journal of Marketing Research*, 442–452.

Fishbein, M. (1979). A theory of reasoned action: some applications and implications.

Forman, C., Ghose, A., and Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291–313.

Gefen, D., Karahanna, E., and Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. *MIS Quarterly*, *27*(1), 51–90.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Ghoshal, A., Menon, S., and Sarkar, S. (2015). Recommendations using information from multiple association rules: A probabilistic approach. *Information Systems Research*, *26*(3), 532–551.

Godes, D., and Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, *23*(4), 545–560.

Godes, D., and Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, *31*(3), 448–473.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press. (http://www.deeplearningbook.org)

Goodhue, D. L. (1995). Understanding user evaluations of information systems. *Management science*, *41*(12), 1827–1844.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2015). Scalable recommendation with poisson factorization. In *Uncertainty in Artificial Intelligence.*

Häubl, G., and Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, *19*(1), 4–21.

Ho, S. Y., and Bodoff, D. (2014). The effects of web personalization on user attitude and behavior: An integration of the elaboration likelihood model and consumer search theory. *MIS Quarterly*, *38*(2), 497–520.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(1), 1303–1347.

Howard, J., and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 328–339).

Hu, Y.-H., and Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management*, *36*(6), 929–944.

Humphreys, A., and Wang, R. J. H. (2017). Automated text analysis for consumer research. *Journal of Consumer Research*, *44*(6), 1274–1306.

Jabr, W., and Zheng, E. (2014). Know yourself and know your enemy: An analysis of firm recommendations and consumer reviews in a competitive environment. *MIS Quarterly*, *38*(3), 635–654.

Jiang, Y., and Guo, H. (2015). Design of consumer review systems and product pricing. *Information Systems Research*, *26*(4), 714–730.

Jo, Y., and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 815–824).

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P., and Welling, M. (2014). Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations.*

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441–504.

Komiak, S. Y., and Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, *30*(4), 941–960.

Komiak, S. Y., and Benbasat, I. (2008). A two-process view of trust and distrust building in recommendation agents: A process-tracing study. *Journal of the Association for Information Systems*, *9*(12), 2.

Kotzias, D., Denil, M., De Freitas, N., and Smyth, P. (2015). From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 597–606).

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behavior. *Information Systems Research*, *13*(2), 205–223.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Kwark, Y., Chen, J., and Raghunathan, S. (2014). Online product reviews: Implications for retailers and competing manufacturers. *Information Systems Research*, *25*(1), 93–110.

Lafferty, J., Mccallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (p. 282-289).

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530–539).

Lee, G., and Lee, W. J. (2009). Psychological reactance to online recommendation services. *Information & Management*, *46*(8), 448–452.

Lee, T. Y., and Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881–894.

Liang, T. P., Lai, H. J., and Ku, Y. C. (2006). Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems*, *23*(3), 45–70.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1–167.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, L., Shang, J., Ren, X., Xu, F. F., Gui, H., Peng, J., and Han, J. (2018). Empower sequence labeling with task-aware neural language model. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

Liu, Q., Gao, Z., Liu, B., and Zhang, Y. (2015). Automated rule selection for aspect extraction in opinion mining. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (pp. 1291–1297).

Liu, Q., Liu, B., Zhang, Y., Kim, D. S., and Gao, Z. (2016). Improving opinion aspect extraction using semantic similarity and aspect associations. In *Proceedings of the thirtieth aaai conference on artificial intelligence* (pp. 2986–2992).

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74–89.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Asso-*

ciation for Computational Linguistics (ACL) System Demonstrations (pp. 55–60). Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (chap. 4). Academic Press.

Miao, Y., Yu, L., and Blunsom, P. (2016). Neural variational inference for text processing. In *International Conference on Machine Learning* (pp. 1727–1736).

Moghaddam, S., and Ester, M. (2011). Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 665–674).

Mudambi, S. M., and Schuff, D. (2010). What makes a helpful review? a study of customer reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185–200.

Naesseth, C., Ruiz, F., Linderman, S., and Blei, D. (2017). Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics* (pp. 489–498).

Nair, V., and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807–814).

Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.

Onishi, H., and Manchanda, P. (2012). Marketing activity, blogging and sales. *International Journal of Research in Marketing*, *29*(3), 221–234.

O'Reilly, C. A. (1982). Variations in decision makers' use of information sources: The impact of quality and accessibility of information. *Academy of Management Journal*, *25*(4), 756–771.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79–86).

Paolacci, G., Chandler, J., Ipeirotis, P. G., et al. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pappas, N., and Popescu-Belis, A. (2014). Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.*

Parsons, J., and Ralph, P. (2014). Generating effective recommendations using viewing-time weighted preferences for attributes. *Journal of the Association for Information Systems*, *15*(8), 484.

Pavlou, P. A., and Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, *17*(4), 392–414.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Petty, R. E., Cacioppo, J. T., and Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, 135–146.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., ... others (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)* (pp. 19–30).

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, *37*(1), 9–27.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015). Deep exponential families. In *Artificial Intelligence and Statistics* (pp. 762–771).

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Conference on empirical methods in natural language processing.*

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Samuelson, P. A. (1948). *Foundations of economic analysis*. Harvard University Press.

Schindler, R. M., and Bickart, B. (2012). Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234–243.

Siering, M., Muntermann, J., and Rajagopalan, B. (2018). Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. *Decision Support Systems*, *108*, 1–12.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., and Roy, P. K. (2017). Predicting the helpfulness of online consumer reviews. *Journal of Business Research*, *70*, 346–355.

Singh, V. P., Hansen, K. T., and Gupta, S. (2005). Modeling preferences for common attributes in multicategory brand choice. *Journal of Marketing Research*, *42*(2), 195–209.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.*

Srivastava, A., and Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Stigler, G. J. (1961). The economics of information. *Journal of political economy*, *69*(3), 213–225.

Tam, K. Y., and Ho, S. Y. (2005). Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information Systems Research*, *16*(3), 271–291.

Tam, K. Y., and Ho, S. Y. (2006). Understanding the impact of web personalization on user information processing and decision outcomes. *MIS Quarterly*, *30*(4), 865–890.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., and Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning* (pp. 190–198).

Timoshenko, A., and Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*.

Tirunillai, S., and Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479.

Titov, I., and McDonald, R. (2008b). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on world wide web* (pp. 111–120).

Titov, I., and McDonald, R. T. (2008a). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Conference 46th Annual Meeting of the Association for Computational Linguistics* (Vol. 8, pp. 308–316).

TripAdvisor. (2015). *Hotel-review datasets.* Retrieved from http://sifaka.cs.uiuc.edu/~wang296/Data/

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417–424).

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wang, H., and Ester, M. (2014). A sentiment-aligned topic model for product aspect rating prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.*

Wang, H., Lu, Y., and Zhai, C. (2010). Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 783–792).

Wang, H., Lu, Y., and Zhai, C. (2011). Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 618–626).

Wang, S., Chen, Z., Fei, G., Liu, B., and Emery, S. (2016). Targeted topic modeling for focused analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1235–1244).

Wang, W., and Benbasat, I. (2016). Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *Journal of Management Information Systems*, *33*(3), 744–775.

WordNet. (2016). *What is wordnet?* Retrieved from http://wordnet.princeton.edu

Wu, Y., and Ester, M. (2015). Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 199–208).

Xiao, B., and Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, *31*(1), 137–209.

Ye, Q., Law, R., and Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*(1), 180–182.

Yelp. (2015). *Yelp dataset challenge.* Retrieved from https://www.yelp.com/dataset_challenge/dataset

Yin, D., Bond, S., and Zhang, H. (2014). Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, *38*(2), 539–560.

Yin, Y., Song, Y., and Zhang, M. (2017). Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2044–2054).

Zhang, Y., and Lin, Z. (2018). Predicting the helpfulness of online product reviews: A multilingual approach. *Electronic Commerce Research and Applications*, *27*, 1–10.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010). Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 56–65).

Zhu, F., and Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*(2), 133–148.