Annals of Operations Research (2024) 339:631–662
https://doi.org/10.1007/s10479-023-05336-z

ORIGINAL RESEARCH

Incorporating topic membership in review rating prediction from unstructured data: a gradient boosting approach

Nan Yang[1] · Nikolaos Korfiatis[1] · Dimitris Zissis[1] · Konstantina Spanaki[2]

Accepted: 14 March 2023 / Published online: 19 June 2023

Abstract
Rating prediction is a crucial element of business analytics as it enables decision-makers to assess service performance based on expressive customer feedback. Enhancing rating score predictions and demand forecasting through incorporating performance features from verbatim text fields, particularly in service quality measurement and customer satisfaction modelling is a key objective in various areas of analytics. A range of methods has been identified in the literature for improving the predictability of customer feedback, including simple bag-of-words-based approaches and advanced supervised machine learning models, which are designed to work with response variables such as Likert-based rating scores. This paper presents a dynamic model that incorporates values from topic membership, an outcome variable from Latent Dirichlet Allocation, with sentiment analysis in an Extreme Gradient Boosting (XGBoost) model used for rating prediction. The results show that, by incorporating features from simple unsupervised machine learning approaches (LDA-based), an 86% prediction accuracy (AUC based) can be achieved on objective rating values. At the same time, a combination of polarity and single-topic membership can yield an even higher accuracy when compared with sentiment text detection tasks both at the document and sentence levels. This study carries significant practical implications since sentiment analysis tasks often require dictionary coverage and domain-specific adjustments depending on the task at hand. To further investigate this result, we used Shapley Additive Values to determine the additive predictability of topic membership values in combination with sentiment-based methods using a dataset of customer reviews from food delivery services.

Keywords Latent dirichlet allocation · Sentiment analysis · Machine learning · Online reviews · XGBoost

✉ Nikolaos Korfiatis
n.korfiatis@uea.ac.uk

[1] Norwich Business School, University of East Anglia, Thomas Paine Study Centre, 1.01C, Norwich Research Park, Norwich NR4 7TJ, UK

[2] Audencia Business School, Nantes, France

1 Introduction

Online reviews are a longstanding topic in literature, and their influence on sales has been well documented (See-To & Ngai, 2018; Liu et al., 2019; Zhang et al., 2020). Customers view reviews as a way to gather information about the quality of products or services (Li et al., 2019; Z. Zhao et al., 2019a, 2019b) as well as embodiments of experience-specific information regarding the quality of products or services. The rating score from customer feedback represents customer overall satisfaction or dissatisfaction directly thus affecting customers' purchase behaviour. As the antecedents of customer rating have been extensively researched, the problem of rating prediction attracts more attention, especially when trying to attribute the rating score to particular aspects of the review or service also known as aspect rating (Korfiatis et al., 2019). The textual content associated with the rating score is provided to justify the latter, thus explaining the underlying rationale, which offers consumers opportunities to express themselves freely rather than feel restricted by pre-defined subareas defined in the user interface (Büschken & Allenby, 2016). Compared with using the characteristics of reviews (i.e., review length) for rating prediction, research concerning the prediction of rating scores from the textual content of customer reviews is attracting more attention, as it tends to carry significant implications for particular service domains.

The review rating score prediction problem is considered to originate from the sentiment classification task of classifying reviews as thumbs up (recommended) or thumbs down (not recommended) (Pang et al., 2002; Qiu et al., 2018). Several studies suggest a high level of consistency between a review's rating score and its textual justification (Hu et al., 2014; Qiu et al., 2018). Even the significant variation of ratings could be explained by customer sentiment statistically (Geetha et al., 2017), the numeric rating scores in customer reviews do not equally represent customer sentiment, as the polarity information appearing in reviews cannot be fully captured by ratings due to the limitations of the rating scale itself (Ghose & Ipeirotis, 2011).

Online reviews are widely adopted for companies to understand how customer perceived the quality of products or services. As demonstrated by Tirunillai and Tellis (2014), multidimensionality exists in quality measurement. Parasuraman et al. (1988) developed SERVQUAL to measure service quality using multiple dimensions including reliability, tangibles, responsiveness, assurance, and empathy. Similarly, there are also latent dimensions within the textual content in online reviews (Tirunillai & Tellis, 2014), The predictive power of these dimensions has been examined by several studies. Korfiatis et al. (2019) illustrated that the extracted dimensions in review text could add predictive accuracy to the overall customer satisfaction. Xu (2020) also revealed the textual factors in review text have an asymmetric influence on customer satisfaction.

The latent dimensions can be discovered using topic models, which has been applied in various business areas to identify different themes from customers' textual feedback. Using topic models, the association and connection between a rating score and review text can be established, which in turn can assist businesses in understanding the reasons driving different levels of customer satisfaction as well as the multidimensionality of the service's outcome (Korfiatis et al., 2019). In topic models, mixed-membership models indicate that a document is considered as a mixture of multiple topics, in which each word belongs to one single topic. Compared with single-membership which allocate each document to only one topic, mixed-membership models allow each document cover multiple topics. we can extract the latent topics (dimensions) as well as how much each document is associated with each latent topic, which is the topic membership. LDA allows us to compute the topic membership and that

contained in unstructured data could increasing the accuracy when predicting rating scores (Korfiatis et al., 2019; Tirunillai & Tellis, 2014).

Therefore, in this study, we aim to evaluate the ability of topic membership to explain and predict customer overall satisfaction as well as exploring how topic membership could be coupled with customer sentiment to enhance the prediction accuracy. Specifically, using machine learning, we examine *how topic models can be used in conjunction with sentiment analysis to quantify customer feedback in cases where domain specific sentiment dictionaries cannot be available.* In addition, given the limitations and the issues with domain-specific sentiment analysis, we evaluate whether *topic modelling can provide a better alternative than sentiment analysis* in review rating prediction. The latter has significant implications, as it removes the necessity for employing domain-specific dictionaries and other approaches that need to be adjusted to the vocabulary of the business domain.

To address these objectives, we perform a large-scale machine learning analysis on a dataset of 1,810,831 customer reviews from 12,153 restaurants on *JustEat*, a popular online food delivery platform in the United Kingdom. Our analysis is multi-faceted. We design and validate two experiments. First, a binary classification task is formed (using a rating score cut-off) to have a robust evaluation of the performance of topic membership as well as its comparison with customer sentiment (in both document-level and sentence-level). We also compare the predictive abilities across topic memberships of all topics. Second, the task of predicting the actual rating score is proposed. We combine each topic membership separately with the polarity score and compare the additive predictability of each topic in predicting the rating score. We perform a post-hoc analysis for robustness by incorporating these two features as covariates using a gradient boosting model using XGBoost—an established machine learning technique. Using the corresponding Shapley additive explanation values (SHAP), we identify the contribution to the prediction of the rating value by each topic/sentiment combination separately.

Our study contributes to the analytics literature by demonstrating how these two different approaches of incorporating features from unstructured data can be used in tandem to predict and explain customer satisfaction. Our findings can lead to faster and more accurate managerial insights for businesses since topic-based rating prediction can uncover multidimensional aspects of service quality that cannot be captured by sentiment analysis. Thus, it can explain customer rating scores and highlight service aspects within customer textual comments to facilitate businesses' understanding of customers' perceived quality towards products or services, which affects customers' purchase decisions (Yeo et al., 2022). Beyond customer reviews from review websites, there are also various sources of customer feedback used in business analytics cases, such as online forums and social media, which do not contain rating scores, or the rating score is incomplete for some of these dimensions. Our study can also extend to these sources.

To this end, the rest of this paper is organised as follows: Sect. 2 reviews the literature on sentiment analysis and topic modelling on customer reviews and how they are applied for rating prediction. Section 3 demonstrates our data sample, models, and metrics for model performance evaluation. Section 4 details the corpus pre-processing procedures and model parameter selection; then, it displays the results for these experiments. Section 5 summarises the analysis and discusses the implications of the results for researchers and practitioners. The study concludes in Sect. 6 with limitations and future research directions.

2 Literature review

2.1 Sentiment analysis

Sentiment analysis is a celebrated computational analysis method in business and management used for detecting customer attitudes and feelings expressed within an unstructured part of customer feedback through natural language processing (NLP) techniques. It is comprised of two main tasks: (a) polarity detection—the identification of whether the text is positive or negative and (b) affect detection—the feelings and emotions expressed in the written communication. Polarity detection is the most common approach applied in sentiment analysis, mainly thanks to the ease of labelling the large textual corpus concerning consumer interaction with company touchpoints (e.g., via social media). In that respect, the literature treats polarity as either an ordered categorical (with the text classified as positive/neutral/negative) or a continuous numerical score of an asymmetric continuum ranging from a normalised negative algebraic value to a positive algebraic value (e.g., -1 to $+1$). Depending on the task at hand, sentiment can be calculated at the document or sentence levels. The latter also produces a confidence band that can produce more reliable prediction outcomes if sentiment is operationalised as input.

Apart from document-level and sentence-level sentiment analysis, aspect-level sentiment analysis is also discussed in the past literature. In many situations, it requires more investigation at the aspect level to identify entities and the associated aspects and sentiments (Do et al., 2019). For instance, companies would like to identify what aspects of their products attract or dissatisfy customers from customer reviews to improve products (Birjali et al., 2021). There could be two types of aspects: explicit aspects and implicit aspects. The former represents those aspects that are directly mentioned in the text, the latter illustrates those aspect terms that don't appear in the text but are implied by other terms (Hu & Liu, 2004). Several Machine learning approaches has been employed to extract the implicit aspects (Bagheri et al., 2013; Quan & Ren, 2014).

Various sentiment analysis techniques are discussed in the literature (Al-Natour & Turetken, 2020; Liu, 2010, 2012; Yadav & Vishwakarma, 2020). These can be summarised into two major types: lexicon-based and machine learning approaches. The lexicon-based sentiment approach uses a bag-of-words model that requires a dictionary consisting of predefined words or phrases assigned by negative or positive values. The other popular approach considers sentiment analysis as a pattern recognition problem and utilises machine learning techniques for classification tasks or predictions (Ghiassi & Lee, 2018). Farkhod et al. (2021) proposed a TDS (Topic Document Sentiment) model, which is an unsupervised machine learning method based on the JST (Joint Sentiment Topic) model and LDA. They used the proposed model to discover the sentiment at the word, document and topic levels. When compared with sentiment terms (usually adjectives) from the lexicon-based approach; machine learning techniques can extract broader and more comprehensive features about several aspects of text, including nouns and verbs expressing descriptions and attitudes towards these objects (Liu, 2012).

In addition, the hybrid approach which combines lexicon-based approach and machine learning approach attracts more attention as it can integrate the advantages of machine learning approach (high accuracy and flexibility) and that from lexicon-based approach (stability) (Birjali et al., 2021). For instance, Marshan et al. (2020) developed a hybrid model combing the lexicon-based and machine learning approaches to detect customer sentiment contained in reviews from an e-commerce platform. Three machine learning approaches are selected

including the Naïve Bayes, KNN (k-nearest neighbours) and SVM (Support Vector Machine). Results showed that the Naïve Bayes had the best performance as a classifier. Elshakankery and Ahmed (2019) proposed a hybrid model HILATSA, representing Hybrid Incremental Learning approach for Arabic Tweets Sentiment Analysis to detect the sentiment in tweets. It is a semi-automatic learning system which will update the lexicon to keep it up to date.

Sentiment analysis for detecting customers' emotions and attitudes, has been widely applied in user-generated content (e.g., online reviews), and several studies have examined the importance of customer sentiment in understanding the relationship with customer ratings, predicting sales, and identifying fraudulent reviews (Y. Zhao et al., 2019a, 2019b; Z. Zhao et al., 2019a, 2019b; Kumar et al., 2022). Innovative applications, combining machine learning and bag-of-words-based approaches, have been applied in practice. Dey et al. (2018) proposed a system that could generate Senti-N-Gram, an n-gram sentiment dictionary, and proposed an algorithm to extract the sentiment scores for n-grams from a random corpus consisting of review text as well as numerical ratings. This approach showed better performance than an existing unigram-based approach (VADER) and another n-gram-based approach (SO-CAL). Recent studies have also applied machine learning approaches to expand dictionary coverage. For instance, Sharma and Dutta (2021) proposed a framework called *SentiDraw*, which calculates the sentiment score for each word from customer reviews based on the rating distribution. Then it was combined with Support Vector Machine (SVM) to achieve better polarity determination.

2.2 Topic models on user-generated content

Online reviews, functioning as the *"voice of the consumer"*, are a form of electronic word-of-mouth (eWOM); these play a critical role in affecting customers' decision-making process, behavioural intention, and product sales performance (Li et al., 2019; Verma & Yadav, 2021). It has been considered an unignorable information source for both customers and sellers, especially the textual content within customer reviews, which includes the textual description of the first-hand usage experiences of previous customers (Guo et al., 2017). The growing popularity of online reviews provides customers with the opportunity to express themselves naturally with unstructured data.

Customers' opinions in online reviews are multidimensional and may reflect different aspects, such as product-specific features or service aspect related evaluations (Büschken & Allenby, 2016; Kim et al., 2020; Mai & Le, 2021). Therefore, to ascertain these latent dimensions from customer reviews, the topic modelling approach is applied widely, as topic models could discover patterns reflecting latent topics within a document from unstructured customer reviews. It assumes that documents consist of a set of topics, and each topic covers a mixture of words (Alghamdi & Alfalqi, 2015). There are a variety of topic models, including Latent Semantic Analysis (LSA) (Landauer et al., 1998), PLSA (Hofmann, 2001), and LDA (Blei et al., 2003). There are many extensions of LDA, including the Correlated Topic Model (CTM) (Blei & Lafferty, 2007) and the Structural Topic Model (STM) (Roberts et al., 2014).

Researchers either adopts existing topic models or proposed new variants of topic models to discover the multidimensionality of customer reviews. The latent dimensions contained in customer reviews are critical since they serve as the foundation for how customers evaluate service, brands and firms, thus affecting new product development or brand positioning. Kwon et al. (2021) employed topic modelling approach and sentiment analysis to online customer review for airlines in order to identify customers' needs. They extracted six dimensions using LDA and identified several words that contained positive and negative emotions

respectively. Tirunillai and Tellis (2014) extended LDA and employed the variant of LDA to customers reviews from five markets and 16 brands. They identified the latent dimensions and ascertained the valence, dynamics and heterogeneity, etc. for strategy analysis. Büschken and Allenby (2016) developed a new model (Sentence-constrained LDA model) based on LDA. They believe that people tend to change their topics across sentences instead of discussing two topics in one sentence. They applied it to two datasets consisting of customer feedback from both restaurant and hotel industry, illustrating the helpfulness of topic modelling approach for unstructured data. Hu et al. (2019) employed STM to customer reviews from hotel industry in order to understand customer dissatisfaction from their complaints. They identified top 10 latent dimensions related with customer dissatisfaction and how dimensions change across hotel grades. Customers of high-grade hotels mainly complained about service issues while that of low-graded hotels are more dissatisfied with facility-related problems.

2.3 Rating prediction

Given the importance of understanding customer satisfaction, rating prediction is a vital task. Several studies have examined the characteristics of online reviews (e.g., review length) to understand customer ratings (Ghasemaghaei et al., 2018; Lai et al., 2021). The information contained in textual comments also plays an important role in understanding customer ratings. Therefore, how to utilise review texts to predict customer rating scores has been a popular topic.

Based on whether the review text is provided in the prediction, the rating prediction task is mainly divided into two categories: (a) personalised rating prediction and (b) review-aware rating prediction. The first focuses on predicting users' rating scores over unrated items using their previous rating behaviours, which is widely explored in the recommendation system field (Cheng et al., 2018; Zhang et al., 2016). Latent factor models, including matrix factorisation, are applied widely and successfully for this type of rating prediction. Customers' textual comments could be corroborated to model user interests and item features (Tan et al., 2016; Zhang et al., 2016) and to improve the accuracy of rating prediction models. The second concentrates on understanding customers' rating scores by discovering valuable information from the provided review text. The direct relationship between sentiment and ratings has been confirmed in previous studies (Geetha et al., 2017; Hu et al., 2014). Table 1 summarises the two categories of rating prediction tasks, which approaches they adopt to extract features from the review text, and what prediction models they adopted.

Büschken and Allenby (2016) employed a variant of LDA to extract latent topics and to predict customer ratings using a dataset of customer reviews from Italian restaurants. It is examined that topic membership could be a meaning device to explain customers' ratings scores. They used a latent cut-point model to examine the relationships between customer satisfaction and the topic membership of 8 topics. Xu (2020) employed LSI to the content of customer reviews from hotel industry and identified 8 positive factors and 17 negative factors. Text regression was conducted using the vector space of each textual reviews to examine how they can affect the overall customer satisfaction. The asymmetric effects were found from the results representing that not all positive textual factors affected customer satisfaction positively. Korfiatis et al. (2019) adopted STM to online reviews from airline passengers and extract latent dimensions of service quality from the textual content. Together with the predefined subcategories by the online platforms, the latent dimensions could add the ability to predict customer overall satisfaction. These studies provide us with another way to predict

Table 1 Overview of existing literature in rating prediction

Type	Task	Approach	Description	Prediction model	Indicative studies
Personalised rating prediction	Extracting contextual features and combining them to rating prediction	Topic model extensions	*These studies were based on topic models, and combined topic modelling approaches with the latent factor models*	Self-proposed model	(Cheng et al., 2018; McAuley & Leskovec, 2013; Zhang & Wang, 2016)
	Extracting complex features and improving rating accuracy	Deep learning	*These studies built deep learning architectures with multiple layers*	Self-proposed model	(Seo et al., 2017; Xing et al., 2019)
Review-aware rating prediction	Exploratory relationship with rating	Sentiment analysis	*Direct impact of sentiment on review ratings and obtained the preliminary relationship between sentiment and ratings*	Fix effects model/linear regression	(Geetha et al., 2017; Hu et al., 2014)
	Rating prediction	Sentiment analysis	*Applied sentiment analysis to review text and adopted the extracted sentiment features for rating prediction*	Ridge regression/linear regression	(Qu et al., 2010; Y. Zhao et al., 2019a, 2019b)
	Rating prediction	Topic models	*Applied LSA to discover positive and negative attributes from reviews and used these attributes as independent variables in text regression to predict overall satisfaction*	Text regression	(Xu, 2020)

Table 1 (continued)

Type	Task	Approach	Description	Prediction model	Indicative studies
	Rating prediction	Both sentiment analysis and topic models	*Applied LDA to exploit several dimensions from review text and combined them with sentiment detected by sentiment analysis to explain the overall satisfaction*	Linear regression/multinomial regression	(Xiang et al., 2017)
	Rating prediction	Self-proposed model	*Added constraints to the LDA model and proposed a sentence-constrained LDA model, and combined it with rating data*	Self-proposed model	(Büschken & Allenby, 2016)

the customer ratings in addition to using the predefined subscales by companies or platforms and proved the ability of topic membership to predict and explain customer ratings.

3 Data and methods

For this study we follow the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is proposed by Wirth and Hipp (2000), aiming to converting business problems into data mining projects which could be carried out and applied regardless of the type of technologies and industries. We illustrate "Business Understanding" and "Data Understanding" in Sects. 3.1 and 3.2. We move to "Data Preparation" in this section and "Modelling" in Sect. 4.

3.1 Dataset

Our data considers textual reviews of UK customers and is collected from *JustEat*, the most popular online food delivery service provider in the UK, with more than 68% of the market share of online orders.[1] Besides, *JustEat* could provide customers with each review including rating score and review text from previous customers who purchased in the specific restaurant while other competitors (e.g., *Deliveroo* and *UberEats*) only display the overall rating score and the number of reviews of the specific restaurant. After customers order and receive the delivered food, they are invited to leave customer reviews describing the entire experience with the food delivery. Potential customers searching for a restaurant can find these reviews on restaurants' pages, which can be of assistance to them. We collect customer reviews written in English and published on their website from January 2016 to November 2021. Generally, there should be a numerical rating score with textual justification in each review. *JustEat* adopts a 6-point scale rating system in which customers give a rating from 1 to 6 stars for three service categories: food quality, delivery time, and restaurant service. The final rating score shown to other consumers when reading these reviews is calculated as the (simple) average of the individual ratings of these three categories.

The textual justification provided by the customer considers all three service categories; therefore, the average rating provides intervals between the minimum and maximum rating scores. To make our analysis more meaningful, we select reviews with textual comments from customers and filter out those that only contain rating scores. Additionally, each review length is constrained in terms of length between 15 and 200 words.[2] In total, our sample contains 1,810,831 customer reviews from 12,153 restaurants. Using the density distribution of the ratings as provided in the dataset, we used the median of the rating scale (3.5 stars) as the boundary (marked with the blue line in Fig. 1) for separating the positive and negative reviews given its even distribution in both classes.

Figure 1 shows the distribution of the rating scores for our sample. This indicates that the percentages of extreme ratings are significantly higher than others. The average rating score is $M = 3.57$ ($SD = 1.82$), which is slightly positive. As to the actual distribution of the scores, the highest percentage occurs in 6-stars ratings, amounting to 20.9%, followed by a 1-star rating, which has the second-highest proportion (15.3%). In total, the proportions

[1] Statista Global Consumer Survey – Brand Report, 2021.

[2] A winsorization procedure was followed for the maximum values considering that any reviews about 200 words were above the 95% quantile of the distribution of review word length.

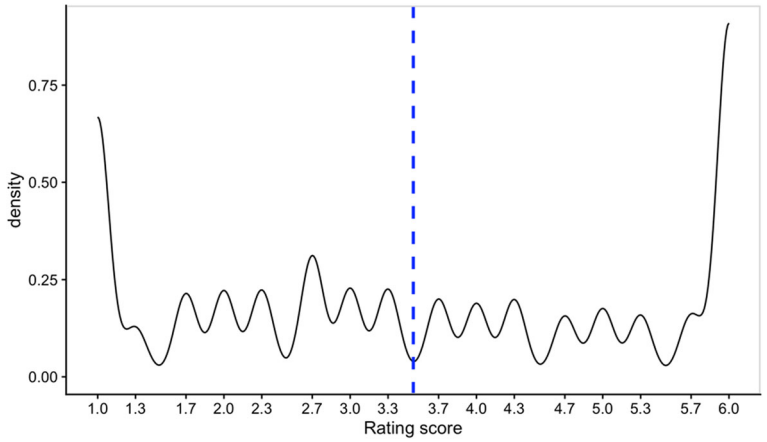

Fig. 1 Distribution of rating scores within the reviews in our sample. The blue dashed line outlines the median rating score

of positive (49.19%) and negative reviews (50.81%) in our sample are similar, which means that our sample is balanced.

3.2 Sentiment analysis

Sentiment analysis is generally measured through polarity, which measures the positive/negative intent expressed in the text. It is generally calculated by various techniques related to the bag-of-words approach; however, other studies in the literature have also applied more complicated models. For polarity calculation, the standard method is to use a lexicon-based approach with domain-specific or general dictionaries (or lexicons). These lexicons can be compiled manually or acquired automatically. The function we adopt to calculate polarity is based on subjectivity lexicons, which contain a list of terms connected with particular emotional states. For instance, the word 'awful' relates to a negative state, while the word 'excellent' is associated with a positive state.

We employ the subjectivity lexicon from Hu and Liu (2004) to calculate polarity, including approximately 6,800 prior-labelled words, which have been identified by benchmarking these terms on a large dataset of online consumer reviews. For identifying polarity in the text, a cluster of terms (x_i^T) that contains four words before and two words after the polarised word has been used to introduce context. Words in the cluster that do not have value are called neutral words and are tagged as x_i^0, which affect word count (n). In addition, there are words that do not have emotion but have an influence on the emotional context, such as valence shifters. Amplifiers (x_i^a)/De-amplifiers (x_i^d)/ are words which can increase/decrease the emotional intent of words, which are given a weight for calculation (Rinker, 2020). The context is defined as follows:

$$x_i^T = \sum\Big((1 + c(x_i^A - x_i^D)\Big) * w(-1) \sum x_i^N,$$

where:

$$x_i^A = \sum\big(w_{neg} * x_i^a\big),\ x_i^D = \max\Big(x_i^{D'},\ -1\Big),$$

$$x_i^{D'} = \sum\left(-w_{neg} * x_i^a + x_i^d\right),$$
$$w_{neg} = \left(\sum x_i^N\right) mod\ 2.$$

The polarity score is calculated as:

$$\delta = \frac{x_i^T}{\sqrt{n}}$$

We constrain the polarity score at $(-1, 1)$ using a transformation formula as follows:

$$\left[\left(1 - \frac{1}{\exp(\delta)}\right) * 2\right] - 1.$$

We use consumers' original review comments before Part-of-Speech tagging and stop words removal because the absence of words will decrease the accuracy by affecting the density of keywords.

3.3 Latent dirichlet allocation (LDA)

The LDA model proposed by Blei et al. (2003) is an unsupervised learning model based on Bayesian inference. Its underlying principle is exchangeability. Compared with latent semantic indexing (LSI) and probabilistic LSI (pLSI) models, LDA considers the exchangeability of both documents and words. LSI applys statistical computations to a large corpus of text to extract and represent the contextual usage meaning of words (Batra & Bawa, 2010). LSI adapts Singular value decomposition (SVD) into the term-document matrix to achieve dimensionality reduction (Zelikovitz & Hirsh, 2001). Deerwester et al. (1990) demonstrated that several basic linguistic notions (e.g., synonymy and polysemy) could be captured by the linear combinations of the tf-idf features, which are derived by LSI. However, the biggest weakness of LSI is the lack of satisfactory statistical foundation. Subsequently, the probabilistic LSI (PLSI), with a solid statistical foundation using a probabilistic method replacing SVD, is proposed by Hofmann (2001) to address the weakness of LSI. It considers each word as a sample from one mixture model, in which the mixture components (multinomial random variables) are considered as "topics". In pLSI, a list of mixing proportions for these mixture components is used to represent each document. However, no probabilistic model at the document-level is not provided, which mean the numbers from each list is not from any generative probabilistic model, leading to overfitting problems seriously (Blei et al., 2003).

LDA is a generative probabilistic model that can deal with sparse vectors of discrete data, including bag-of-words in text data and image features. For text data, the core assumption is that each document is considered a random mixture of latent topics, while each topic is represented by a multinominal distribution over words. LDA is based on the assumption that the author of each document would have the same probability to use same words when writing the same "*topic*". LDA is a generative probabilistic model, which simulates the process of an author producing a document. In this process, the probability of writing a word is related with the topic that are written about. However, if two authors have different words to write for the same document, the distribution of words might change to an unrelated topic by making inaccurate inferences.

Every document is created by a list of hypothetical and unobservable *'topics'*. Each document is assumed to be presented by a mixture of topics that reflect distributions sharing common Dirichlet priors. In a single document, the probability of each topic is between 0

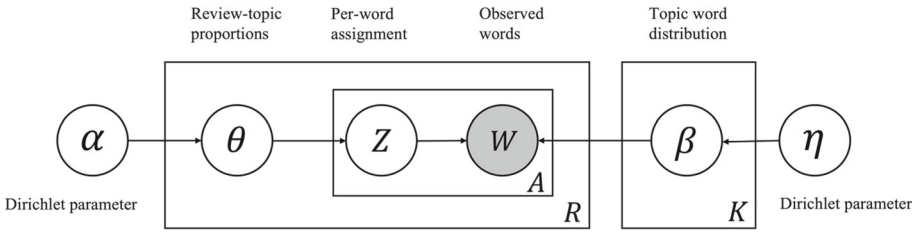

Fig. 2 LDA model process using plate notation (Adopted by Blei et al., 2003)

and 1, with the sum of them amounting to 1. The extent to which documents are associated with topics is considered document-topic proportions, also known as topic membership. The latent topics are considered as the distributions over a fixed vocabulary in which each word has a possibility belonging to each latent topic.

In this study, each review is a single document. We index each review as $r \in (1, 2 \ldots, R)$. K presents the number of topics, which is the primary input variable. The generation process is summarised as follows:

For each topic $k \in (1, 2 \ldots, K)$, draw a Dirichlet distribution over the vocabulary V, $\beta_k \sim Dir(\eta)$.

For each review, r, choose $\theta_r \sim Dir(\alpha)$.

i. For each word w_a, from review r, a topic assignment is drawn from a multinominal distribution over θ_r, $z_{r,a} \sim Mult(\theta_r)$, where $z_{r,a}$ represents the word-specific topic assignment.
ii. The observed word,$w_{r,a}$ is drawn from $Mult(\beta_{z,r,a})$, where $w_{r,a} \in (1, 2 \ldots V)$.

The joint distribution of all unobserved variables and observed variables is expressed as follows:

$$p(\beta_K,\ \theta_R,\ z_R,\ w_R|\alpha,\ \eta) = \prod_{k=1}^{K} p(\beta_K|\eta) \prod_{r=1}^{R} p(\theta_r|\alpha) \prod_{a=1}^{N} p\left(z_{r,a}|\theta_r\right) p\left(w_{r,a}|z_{r,a},\ \beta_{r,k}\right)$$

Figure 2 depicts the graphical model of the LDA process in plate notation. The shaded nodes present the only observed variables $w_{r,a}$, which represents the a th word in review r. A is the total number of words in each review. Each review could be represented as a mixture of topics. θ denotes the review-topic distribution, which indicates how much each review is related to topics. β denotes the per-review topic–word distributions. The problem of sparsity caused by the large vocabulary occurs in many document corpora. The *smooth* method is usually adopted to avoid assigning zero probability to words that are from new documents but don't appear in documents from training corpus (Blei et al., 2003). Instead of the commonly used method-Laplace smoothing, LDA places the Dirichlet priors on the multinomial parameters. Therefore, α and η, as two Dirichlet parameters, denote the smoothing of topics with reviews and words within topics, respectively (Syed & Spruit, 2017).

The topic-word distributions and the coefficients for documents cannot be observed and are estimated by a learning algorithm, such as expectation propagation (a higher-order variational algorithm) and the Markov chain Monte Carlo algorithm (Griffiths & Steyvers, 2002; Minka & Lafferty, 2002; Porteous et al., 2008). For model inference and parameter estimation, we adopted Gibb's sampling to compute the approximations to the posterior distribution of the hidden variables in the model, which is the core inferential problem in LDA. Compared

with the convexity-based variational approach introduced by Blei et al. (2003), Gibb's sampling could achieve higher accuracy by approaching the asymptotically correct distribution (Porteous et al., 2008).

3.4 Extreme gradient boosting (XGBoost)

To examine the ability of two approaches in predicting rating scores, we conduct a set of two-stage experiments: binary classification and rating prediction. Classification is used to identify which category the new observation belongs to, while prediction involves making future estimations based on current data behaviour patterns (Brintrup, 2021). Extreme Gradient Boosting (XGBoost) is a highly effective scalable tree-boosting system. It has achieved state-of-the-art results on a wide range of machine learning challenges because of its effectiveness, flexibility, and portability (Chen & Guestrin, 2016).

Giannakas et al. (2021) compared the ability of XGBoost with a 4-hidden-layers Deep Neural Network (DNN) when making prediction of the team performance. The results revealed that both the learning accuracy and prediction accuracy of XGBoost are higher than DNN. Wu et al. (2021) applied five different datasets to examine the performances of XGBoost and Multiple-layer Perceptron Neural Network for binary classification tasks. The results demonstrated that XGBoost performed generally better than the neural network and significantly better when the overlapped samples increased. Khanam et al. (2021) evaluated the performance 7 algorithms including Logistic Regression, XGBoost, KNN, Naïve Bayes, Decision, SVM and Random Forests when performing classification task for fake news detection. It is examined that XGBoost depicted the highest accuracy than other algorithms. Rao et al. (2021) also compared the performance of several algorithms as classifiers including XGBoost, Logistic Regression, Random Forest, Decision Tree, Multinomial Naïve Bayes and Bernoulli Naïve Bayes to perform the binary classification task of detecting fake news. They demonstrated that XGBoost could provide excellent mix of prediction and processing speed simultaneously. After fine-tuning hyperparameters, it could achieve the highest accuracy than other methods. Apart from classification, another study from Yan et al. (2022) examined the power of XGBoost to make predictions in health field. They compared XGBoost with multivariate logistic regression model and found that the former performed better in predicting the risk of death with one specific disease. Due to the better performance of XGBoost compared with other algorithms, we believe that XGBoost is an appropriate approach for classification and prediction tasks.

The gradient boosting approach is described as follows: Assume a dataset $D = \{(x_i, y_i) : i = 1, \ldots, n, x_i \in R^m, y_i \in R^n\}$, with n instances and m features.

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \;\; f_k \in F$$

where K represents the total number of trees, $f_k(x_i)$ is the predicted value of i-th sample in the k-th tree,F is the function space consisting of all CARTs (regression or classification trees in XGBoost), and $\widehat{y_i}$ is the predicted value, for instance, i.

The set of functions $f(k)$ could be learned by minimising the objective function, which consists of training loss $L(\theta)$ and regularisation term $\Omega(\theta)$.

$$O(\theta) = L(\theta) + \Omega(\theta) = \sum_{i}^{n} l(y_i, \; \widehat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

where l is the training losing function, measuring the difference between the predicted value $\widehat{y_i}$ and the observed value y_i The regularisation term Ω could control the model complexity to avoid overfitting.

The tree model could be trained using an additive strategy.

$$\widehat{y_i}^t = \widehat{y_i}^{t-1} + f_t(x_i)$$

Therefore, the objective function at step t is changed as follows:

$$O^{(t)} = \sum_{i=1}^{n} l\big(y_i,\ \widehat{y_i}^{t-1} + f_t(x_i)\big) + \Omega(f_t) + constant$$

By using the second-order Taylor expansion, we simplify the equation as below:

$$O^{(t)} = \sum_{i=1}^{n} [\left(l\big(y_i,\ \widehat{y_i}^{t-1}\big) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + constant$$

where g_i and h_i are represented as follows:

$$g_i = \partial_{\widehat{y_i}^{t-1}} l\big(y_i,\ \widehat{y_i}^{t-1}\big)$$

$$h_i = \partial^2_{\widehat{y_i}^{t-1}} l\big(y_i,\ \widehat{y_i}^{t-1}\big)$$

In XGBoost, the complexity is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

where γ and λ are regularisation parameters, T represents the number of leaves and w are scores on leaves. By defining $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ and expanding the regularisation term, the objective function is re-formulated as:

$$O^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}\big(H_j + \lambda\big) w_j^2] + \gamma T$$

The best w_j^* and the best corresponding value could be computed as

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

$$O_j^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \lambda T$$

Given that enumerating all possible trees is not intractable, the tree is optimised on one level at a time by splitting leaves and producing a gain score:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

3.5 3.5. Model performance measurement metrics

In our study, we use two types of metrics to compare the performance of the classification and prediction models. For the classification, a standard approach utilising a confusion matrix is used to represent the dispositions of the test dataset in a 2×2 setting (true positive, true negative, false negative, false positive).

The true positive rate, also known as recall or sensitivity, is calculated as:

$$Sensitivity = \frac{TP}{TP + FN}$$

The false negative rate, also known as specificity, is estimated as:

$$Specificity = \frac{TN}{TN + FP}$$

Precision is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

An equal weighted combination of these two metrics can be reflected on the F1-score, which can be calculated as:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The false positive rate (FPR) is equivalent to $1 - Specificity$. The ROC curve is a two-dimensional plot that shows sensitivity on the y-axis and $1 - Specificity$ (FPR) on the x-axis. The perfect one is located at point (0, 1). The ROC curve starts from point (0, 0) and ends at point (1, 1). AUC is a method to compare classifiers and is calculated as the area under the curve, which is between 0 and 1. The model with higher AUC performs better in classification than others, as it is known that *"the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance"* (Fawcett, 2006, p. 868).

In addition to the binary classification, we would perform regression using XGBoost. Thus, to determine the effectiveness of our model, two metrics will be calculated: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE represents the average of the difference between the predicted value and the original value, which can be calculated by the formula below. A smaller MAE indicates a better model.

$$Mean\ Absolute\ Error = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}|$$

RMSE is popularly used in the literature, which represents the square root of mean squared error, between the actual and predicted rating score (y) as follows:

$$Root\ Mean\ Squared\ Error = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}$$

All metrics are estimated and used in evaluating the performance of each model in the analysis.

3.6 Feature selection

Unlike simple models (e.g., linear regression), more complex predictive models (e.g., deep learning and tree-based models) are complicated to interpret. Shapley Additive explanations (SHAP) values, proposed by Lundberg and Lee (2017), could better interpret black-box models by computing Shapley values from coalitional game theory. The Shapley value is an explanation method based on solid theory, in which four axioms (efficiency, symmetry, dummy, and additivity) provide a reasonable foundation. It represents how much a feature contributes to the prediction for each instance compared with the average prediction of the trained model (Molnar, 2020). Inspired by cooperative game theory, the Shapley value is a method of fairly distributing pay-outs to players according to their contribution. In this study's circumstances, the 'players' represent the feature values, and the 'game' signifies the prediction task.

SHAP could explain the Shapley values as a linear model specified as:

$$g(z) = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j z_j$$

where M is the maximum number of simplified input features. $z \in \{0, 1\}^M$. When calculating the Shapley value, the value of z represents the status of the presence (used or not used) of the corresponding feature in prediction. $\varnothing_j$ is the Shapley value (the attribution of feature j). The Shapley value of feature j can be calculated as follows:

$$\emptyset(j) = \sum_{s \subseteq \{1,\ldots,M\} \setminus \{j\}}^{M} \frac{|S|!(M - |S| - 1)!}{M!} (v(S \cup \{j\}) - v(S))$$

where S represents one subset of the simplified features included in the model. $v(S)$ is the total value for S. The marginal contribution of feature j is calculated as $v(S \cup \{j\}) - v(S)$.

SHAP could be a powerful method to interpret results from tree-based machine learning models (e.g., random forest and gradient boosted trees). In this study, SHAP demonstrates the feature importance by examining its marginal contribution to the model output, which provides local explanation and consistency globally.

4 Results

4.1 Baseline sentiment calculation

For each review, we calculate the sentiment polarity score (document-level and sentence-level polarity) using the original textual comment at the review level and provide a decimal between − 1 and 1. Figure 3 provides the distribution of document-level polarity scores in our sample. The biggest peak of the curve is in the middle. The average polarity score is 0.05, and the standard deviation equals 0.28. The most frequently occurring polarity values are clustered near the middle. The extreme polarity scores (close to 1.0 and − 1.0) occur the least frequently, which is quite different from the distribution of customer rating scores (Fig. 1), as extreme rating scores show the most frequent occurrence.

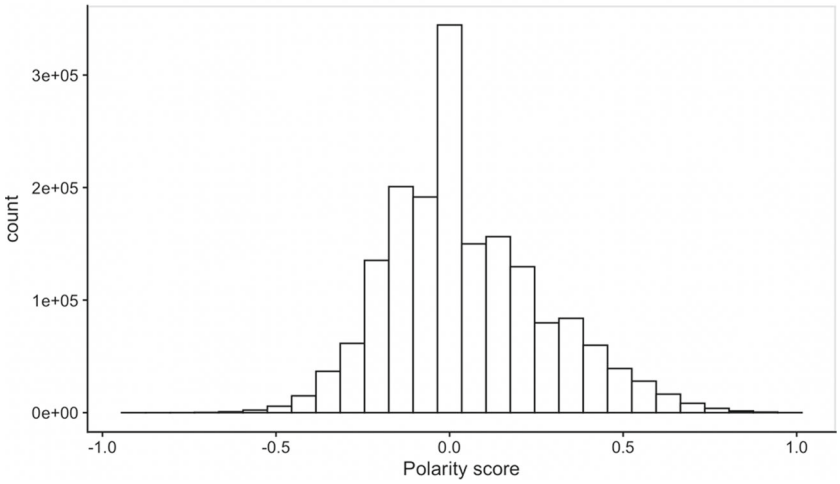

Fig. 3 Polarity score distribution for our sample

4.2 Corpus pre-processing for topic modelling

The textual content from customer reviews is pre-processed by following the standard procedural remedies, including (a) word tokenization (breaking sentences into a set of tokens), (b) exclusion of numbers and punctuations, (c) elimination of stop words, which includes both language stop words removal using the SMART stop word list and context-specific stop words exclusion, such as food vocabulary and restaurant brand names, and (d) selecting only adjectives, nouns, and adverbs from remaining words, since these words contain relevant information about products and product quality (Tirunillai & Tellis, 2014). This step is implemented by utilising part-of-speech (POS) tagging to keep the parts of speech that are meaningful as well as lemmatization, which derives the base forms of the words. (e) For low-frequency words (frequency of occurrence is less than 2% of the total number of reviews), a pruning procedure is followed, which reduces the number of reviews to *1,700,131*. The low-frequency words which convey highly specific semantic information are considered as the weak features in the corpus (Leeman, 2007). We followed the procedure of removing low-frequency words based on the study of Griffiths and Steyvers (2004).

4.3 LDA model estimation and hyperparameter tuning

After transforming textual data into a document-term matrix, we estimate the topic models and use a heuristic approach to evaluate the hyperparameter values that provide an ideal solution using the current parameter set. As shown in Fig. 2, there are two hyperparameters, α and η, which are two smoothing parameters controlling the sparseness of Dirichlet distribution. These two values and the number of topics K are required to be inputted for the LDA process. To determine the best number of K, many researchers have adopted trial and error procedures. A set of models was estimated with various values of K and the model producing the most meaningful topics is selected (Bastani et al., 2019; Blei, 2012). Based on the intrinsic nature of reviews from the OFD platform, we could infer that reviews are homogeneous and concentrated on only a few themes (e.g., food quality, delivery speed, driver's attitude). As

JustEat adopted a 6-star rating system, we estimate 13 LDA models with different values of K starting from 6 to 18.

Several researchers seek to find the best number by calculating the *'perplexity'* of the held-out test set, which is one intrinsic evaluation metric for language model evaluation. Perplexity algebraically equals the inverse of the geometric mean per-word likelihood (Blei et al., 2003), which means that the lower perplexity indicates that the model predicts better for new test samples. However, there is a distinct drawback of using perplexity to evaluate the quality of the LDA model. The perplexity decreases as the number of topics increases (Koltcov et al., 2014). Chang et al. (2009) illustrated that producing ever finer partitions as the number of topics grew could make the model less helpful and reduce topic interpretability.

Therefore, to find the best number of K, two metrics are calculated, proposed by Cao et al. (2009) and Deveaud et al. (2014), to compare 13 LDA models. Cao et al. (2009) found that the best K is not only correlated with the size of the dataset but is also influenced by the inherent correlations within the corpus. They considered each topic as a semantic cluster, in which the similarity of each word is as small as possible, while the similarities among topics are expected to be large. Similar with the idea of clustering based on density, they aim to achieve a large similarity within the topic for more explicit semantic meaning while a small similarity among topics showing a stable topic structure. The procedures are as follows: First, the initial LDA model is estimated given an arbitrary K value. Second, they calculate the average cosine distance of the model, the model's cardinality, and all topics' density. Third, based on the cardinality, they re-estimate the LDA model and initialise sufficient statistics. If the direction of convergence is negative, topics with high densities will be applied as reference samples. Otherwise, the seeded method will be adopted to initialise it. Then, repeat the second and third steps until the model's average cosine distance and cardinality converge.

In addition, Deveaud et al. (2014) proposed a simple heuristic approach to find the best number of topics when the information diverges between all pairs within the LDA model. Rather than the non-symmetric measure (Kullback–Leibler divergence), the symmetrised version of Jensen-Shannon divergence is applied. Figure 4 depicts the performance of different LDA models using these types of metrics for the values of K (x-axis). The model achieves the best performance when the upper metric has the minimum value or the lower one is maximised. Therefore, we select 15 as the optimal value of K.

4.4 Topic identification

Table 2 provides the K = 15 topic solution for the review corpus, which is the optimised solution after topic number selection, as previously discussed. The loading words associated with each topic can be used to understand the main concerns and preferences of customers in relation to the food delivery service being reviewed. The topic solution covers 15 topics' top 7 loading words separately produced using the standard topic word probability (β) from the LDA estimation process. Several topics show positive intention, such as Topics #1, #2, #3, and #5, while some topics are more negative (i.e., Topics #11, #12, and #13). For instance, Topic #12 mainly talks about the delivery service from drivers about locating their address. Topic #13 concentrates on late deliveries and long waiting times. Topic #3 focuses on customers' praise and subjectively positive descriptions of their takeaways.

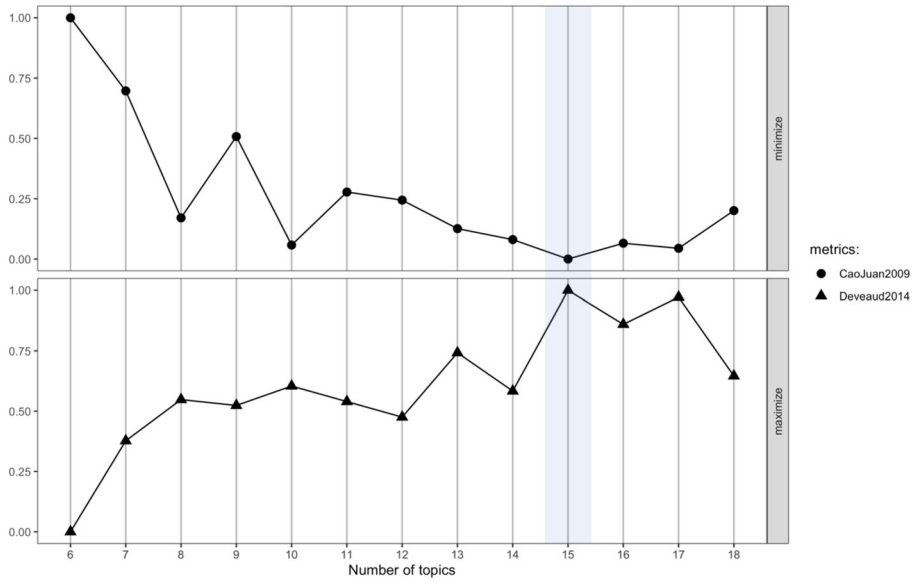

Fig. 4 Selection of the number of topics (K) for identifying the topic solution. Optimal K is identified by the shaded area

Table 2 The 15 topics and their top 7 loading words in the topic solution

Topic #	Top 7 loading words
Topic 1	Food, cold, stone, longer, late, delivery, driver
Topic 2	Hot, nice, food, lovely, fresh, tasty, again
Topic 3	Best, place, amazing, delicious, guy, takeaway, excellent
Topic 4	Service, back, customer, phone, poor, bad, problem
Topic 5	Food, great, always, early, delivery, fast, home
Topic 6	Good, quality, portion, large, small, price, worth
Topic 7	Meal, only, happy, extra, box, thing, instead
Topic 8	Order, wrong, right, issue, correct, number, store
Topic 9	Not, more, disappointed, taste, lot, flavour, same
Topic 10	Food, again, warm, free, once, hungry, barely
Topic 11	Never, again, money, ever, dry, soggy, hard
Topic 12	Delivery, driver, where, door, address, house, man
Topic 13	Late, hour, minute, min, half, way, later
Topic 14	Time, first, last, long, few, second, next
Topic15	Drink, item, missing, bag, refund, order, full

4.5 Incorporating topic membership in sentiment text detection

As shown in the literature review, sentiment and topic membership could be considered as two devices to explain and predict customer satisfaction. We would like to examine how sentiment and topic membership can predict customer satisfaction empirically. As mentioned in Sect. 3.1, we consider the mid-point of the rating scale (3.5 stars) as the boundary to separate negative and positive reviews. Based on its rating score, each review in our sample is classified into two classes: *positive* and *negative*. Therefore, by classifying customer reviews into positive and negative, the question is transformed to a binary classification task. More specially, we would also examine the different ability of document-level sentiment and sentence-level sentiment in this classification task. Therefore, we constructed three models (Table 3) with all target variables being rating (positive/negative) for the classification task. Model A and Model B include the calculated polarity score (document-level and sentence-level separately) of each review as independent variables to predict the class. Model C adopted the topic membership of 15 topics from the topic solution of the LDA process as the predictors. The in-sample validation split was 80% for training and 20% for testing with additional folds selected for confidence interval estimation. For both classification tasks, the XGBoost algorithm is followed. For hyperparameter selection, we limit the maximum depth of the tree to 2 and the maximum rounds of boosting iterations to 100 to obtain the best outcome.

AUC scores are calculated using k-folding, and ROC curves are graphically presented in Fig. 5. As presented in Fig. 5, the dashed curve represents the ROC curves of Model A and Model B (using document-level polarity and sentence-level polarity separately), whose AUC scores are 0.827 (95% CI: 0.825–0.828) and 0.838 (95% CI: 0.837–0.840), respectively showing relatively good classification, as they are both larger than 0.8. The sentence-level polarity score has better performance than the document-level polarity score in classifying the two rating classes. The solid line is the ROC curve of Model C using topic membership (15 topics), demonstrating better performance than using polarity; the curve close to the upper left corner indicates higher accuracy. The AUC score is 0.860 (95% CI: 0.8587–0.8611), which represents a better ability to separate positive and negative classes. Only comparing the two models with polarity, solely using topic membership on its own could increase the predictive accuracy.

As mentioned, topic membership, including all topics, performs better than sentiment in helping to classify the positive and negative reviews. These 15 latent topics extracted through the LDA process indicate various dimensions customers pay attention to contained in customer reviews. Some of them might be emotional and highly related to customers' ratings, while some might be more realistic. To examine the individual contribution of each topic to the predictive accuracy of binary classification, we construct and perform 15 models

Table 3 Three models' AUC Values for the classification task

Target variable	Polarity (Document)	Polarity (Sentence)	Topic membership (15 topics)
Rating (positive/negative)			
Model A	0.827		
Model B		0.838	
Model C			**0.860**

Highest scoring model highlighted in bold

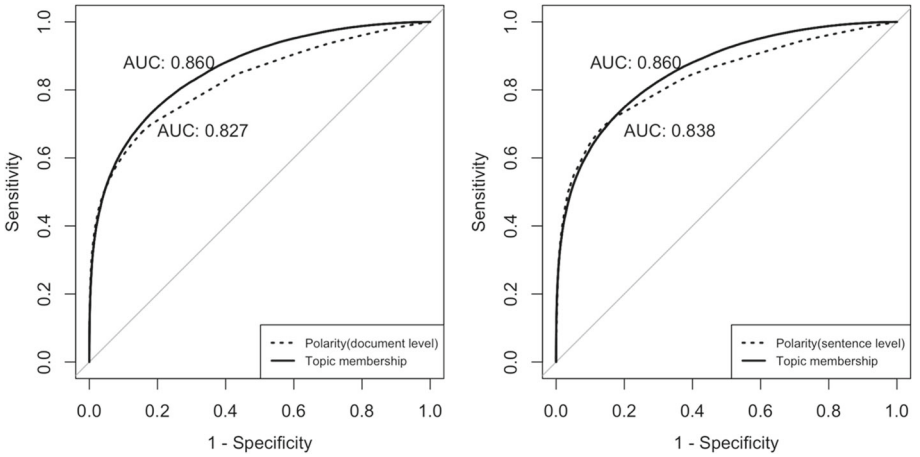

Fig. 5 AUC comparison (left-hand side) between Model A and Model C as well as AUC comparison (right-hand side) between Model B and Model C

(with the same hyperparameters as models in the previous experiment) to classify positive and negative classes and include the topic membership for each topic as predictors. Table 4 demonstrates the AUC scores for the 15 models. Several models (Models #4, #7, #9, #10, #12, and #14) have relatively low AUC scores (close to 0.5) and do not help much in predicting customer attitudes. The remaining models have higher predictive performance for classification in some way, whose AUC scores are higher than 0.6. Among them, Model 3 has the highest AUC score (0.706) and presents the best performance in the binary classification task. The dimension that Topic #3 mainly talks about indicates the strongest relationship with customers' attitudes (positive or negative).

We already examined the model with topic membership and model with sentiment only separately and proved that topic membership with all topics included could perform better than polarity only in classifying positive and negative classes. However, the combined predictive power of sentiment and topic membership has not yet been examined. As Topic #3 contributes most to the predictive accuracy of classification, we construct Model D and Model E with the integration of Topic #3 and polarity (two levels) as predictors for the classification as shown in Table 5. The former includes document-level polarity score and topic membership (only Topic #3) while the latter includes sentence-level polarity score and topic membership (only Topic #3) as predictors.

Models were performed with the same hyperparameters, and their ROC curves and AUC scores are displayed and compared with the Model A and Model B, as shown in Fig. 6. The comparison (the left-hand side) of Model A and Model D showed that Topic #3 added as a new predictor together with document-level polarity could increase the AUC score to 0.842 (95% CI: 0.841–0.843). The comparison (the right-hand side) of Model B and Model E revealed that Topic #3, together with the sentence-level polarity, could increase the AUC score to 0.852 (95% CI: 0.851–0.854). It demonstrates that topic membership possesses the additional power to help sentiment when predicting customer attitudes.

The parameters that decide the model architecture are hyperparameters. We could find the ideal hyperparameters and improve our predictive accuracy through hyperparameter tuning. After tuning hyperparameters, we find a relatively better list of hyperparameters with a

Table 4 AUC Comparison among 15 topics' memberships in the classification task

Predictor	Topic1	Topic2	Topic3	Topic 4	Topic 5	Topic 6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15
Model 1	0.621														
Model 2		0.677													
Model 3			**0.706**												
Model 4				0.571											
Model 5					0.637										
Model 6						0.612									
Model 7							0.546								
Model 8								0.605							
Model 9									0.548						
Model 10										0.539					
Model 11											0.618				
Model 12												0.542			
Model 13													0.649		
Model 14														0.551	
Model 15															0.626

Highest scoring model highlighted in bold

Table 5 Two models construction with combination of sentiment and topic (#3) membership

Target variable	Topic membership (only topic 3)	Polarity (Document level)	Polarity (Sentence level)	AUC
Rating (positive/negative)				
Model D	•	•		0.842
Model E	•		•	0.852

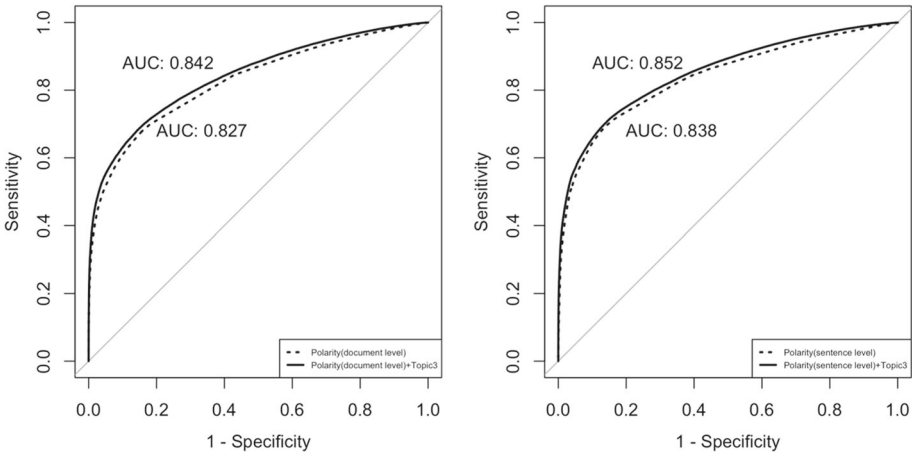

Fig. 6 AUC comparison (left-hand side) between Model A and Model D as well as AUC comparison (right-hand side) between Model B and Model E

learning rate of 0.3, gamma of 0.2, maximum depth of a tree as 7, and the minimum sum of instance weight as 5 in a child, and the subsample ratio as 0.8. Finally, as Fig. 7 shows, we improve the AUC scores of models with polarity (both document-level and sentence-level) and Topic #3 as predictors to 0.845 (95% CI: 0.843–0.846) and 0.856 (95% CI: 0.854–0.857), respectively.

4.6 Rating score prediction

To examine the ability of polarity and topic membership to predict the exact rating score, we construct the baseline model, which includes the (sentence-level) polarity score of each review as the only predictor. The sentence-level polarity will be adopted in the rating score prediction, considering that it is more accurate than document-level performance from the results in the previous section. By combining each topic membership separately with polarity score, 15 models are formed and trained using the same training dataset used in the binary classification task and performed using XGBoost. Considering that this task is an actual prediction task, MAE, and RMSE models are used for evaluation.

We calculate MAEs and RMSEs for 15 models as well as the baseline model. Figure 8 displays the relative difference of MAE and RMSE for 15 models compared with the baseline model and sorts from the highest change to the lowest change. All topic membership could

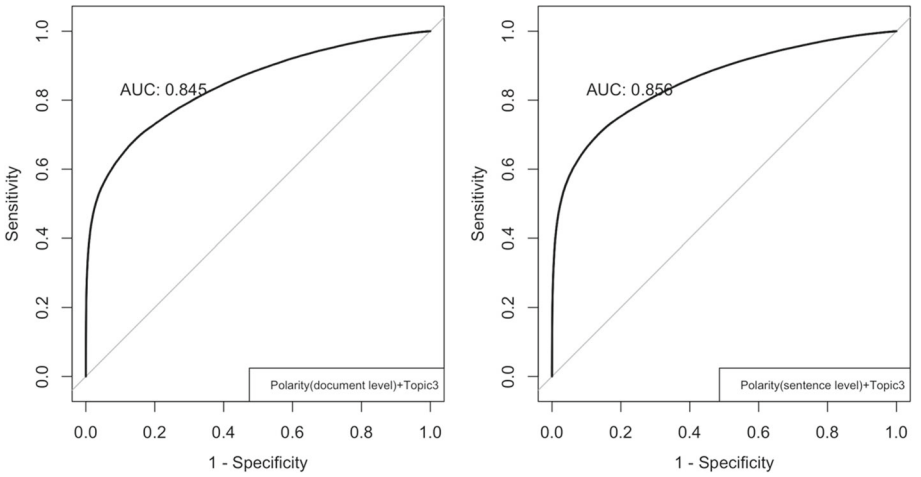

Fig. 7 ROC curves for Model D and Model E after tuning the hyperparameters of XGBoost

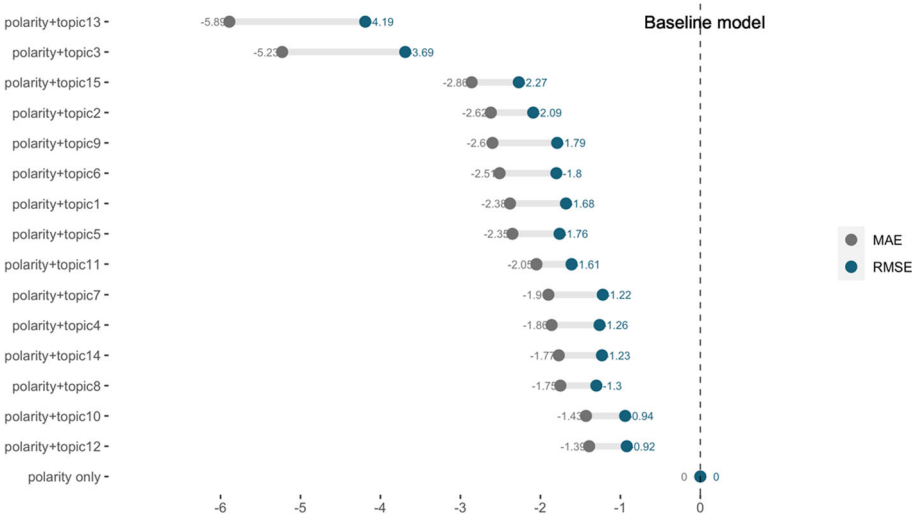

Fig. 8 Relative difference of MAE and RMSE for 16 models compared with the baseline model

decrease the error compared with the baseline model, which indicates that the inclusion of even one topic as a covariate could increase the accuracy of the prediction task. There are two distinct variables (Topics #13 and #3) that can dramatically decrease the MAE and RMSE.

4.7 Shapley additive explanations (SHAP) for each feature

While Topic #13 and Topic #3 perform best in reducing the model error, other topics (i.e., Topics #15 and #2) also show improvement compared with the baseline model. To get more specific and direct comprehension of how much each topic can contribute to the prediction

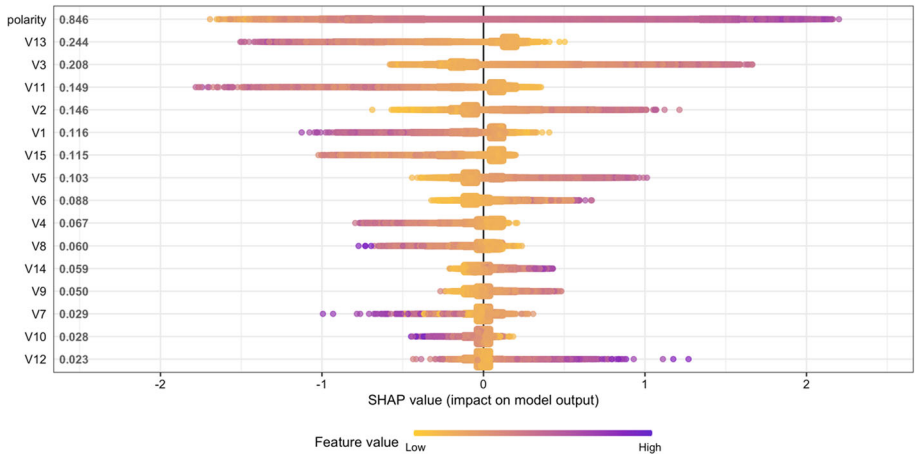

Fig. 9 SHAP summary plot

of customer ratings, we construct a model including polarity score (sentence-level) and proportions of 15 topics as covariates to predict customer ratings. The accuracy is improved substantially (MAE = 0.869, RMSE = 1.137) by including all topics. The contribution of each feature to the target value is represented by Shapley additive explanations for feature importance, which is calculated as the average of the absolute Shapley values for each feature across the dataset. SHAP values can be used to understand the relative importance of different features in a model. These values are calculated by comparing the model's output with the expected value of the model's output over the entire distribution, taking into account the dependencies between features. By using SHAP values, we can identify the most important features in a model and understand how they contribute to its predictions.

The summary plot (Fig. 9) displays global feature importance, as well as feature effects. All variables are sorted by decreasing feature importance along the y-axis, with their corresponding value next to them. The polarity score is the most dominant feature, and Topic #13 is the second most important feature, followed by Topic #3, while Topic #12 contributes the least to the predicted values. Each dot in this plot shows the Shapley value of an instance for each feature, whose horizontal location displays its Shapley value, and vertical location is determined by the specific feature. The gradient colour demonstrates the original value for that variable from low to high. Polarity affects the target variable positively, as high polarity scores could increase the predicted customer ratings. Topic #13's membership is negatively associated with the target value as the predicted rating score will decrease while the proportion of Topic #13 within a review increases.

Figure 10 displays the dependence plot for Topic #13 and its interaction with the polarity value. Each dot represents an instance with its proportion within a single review on the x-axis and its corresponding Shapley value on the y-axis. The gradient colour shows its polarity value. A small number of dots with proportions lower than 0.067, have positive SHAP values, indicating an increased prediction value. In contrast, a great many instances have a higher topic proportion between 0.067 and 0.096. Their corresponding SHAP values are lower than 0, meaning that they decrease the predicted value. More explicitly, for these dots whose x-axis is between 0.067 and 0.096, as the proportion of Topic #3 increases, their negative influence on the predicted rating score will be stronger.

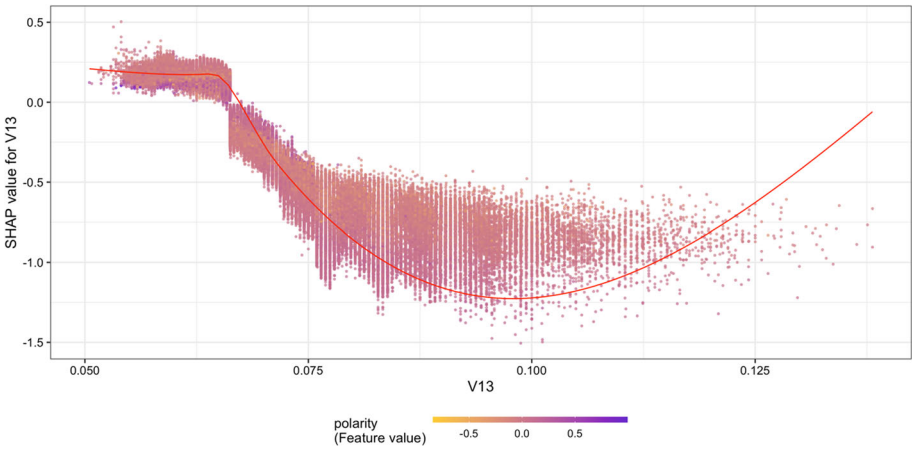

Fig. 10 SHAP dependence plot for Topic #13 membership and its interaction visualisation with polarity score

5 Discussion and implications

5.1 Discussion

Through the robust experiment (binary classification), the results show that both document (review)-level polarity score and sentence-level polarity score perform well in classifying a review positive or negative with AUC scores—with both being higher than 0.8 even they are included as the only covariate in two models, respectively. Compared with document-level polarity, sentence-level polarity performs better in the classification task, with a higher AUC score (0.838) than the other AUC score (0.827). Polarity within customer textual content could excellently predict customer satisfaction, while sentence-level polarity has a better ability for the prediction, which identifies the impact of different granularity levels. It confirms the strong ability of sentiment to explain customer ratings, consistent with Chatterjee et al. (2021) and Zhao et al., (2019a, 2019b).

However, the topic memberships of 15 latent topics extracted from review text using a topic modelling approach (LDA) perform better in the classification task, with an AUC score of 0.860. Topic membership (all topics included) has higher accuracy than the classification task's polarity score (both document-level and sentence level). It represents the multidimensionality exists in customer reviews in which they discuss their opinions towards the food and deliver service from various aspects. The multidimensionality is consistent with other studies which focused on customer reviews from other hospitality industries such as airlines, restaurants, and hotels (Büschken & Allenby, 2016; Xu, 2020). In addition, the multidimensionality not only stands for various entities (e.g., food quality and delivery service), but also demonstrates customers' dialectical in textual content. Customers might describe their experiences dialectically, such as two-sided reviews (Wang et al., 2022). That may explain the stronger ability of topic membership to predict ratings as both document-level and sentence-level sentiment could not capture the various dimensions (Birjali et al., 2021).

By examining each topic membership to the classification task separately, the one with the highest AUC score is selected and collaborated with the polarity score (two different levels), improving the AUC scores to 0.845 and 0.856, respectively, after hyperparameter tuning.

The robust check could prove that the features generated from the textual content could be combined with the sentiment to achieve higher accuracy. It also reveals the heterogeneity of the latent dimensions (Büschken & Allenby, 2016; Hu et al., 2019), which don't equally contribute to customer's overall ratings.

Therefore, we included polarity (sentence-level) and each topic membership one by one as covariates and constructed 15 regression models. Compared with the baseline model (with only sentence-level polarity score), the results indicate that whichever topic membership could improve the model performance. Among them, two topics (Topic #13and #3) could add the most accuracy to polarity in predicting rating scores since the MAE and RMSE of that model are decreased most compared to the baseline model. The information captured by the two topics could add more predictive accuracy to sentiment for rating prediction, which is also proved by the SHAP feature importance when we include all topics' membership and polarity (sentence-level) into the prediction.

5.2 Theoretical implications

Many studies have examined the power of sentiment and latent dimensions within review text to explain and predict customer overall satisfaction (Cheng et al., 2018; Xing et al., 2019; Zhao et al., 2019a, 2019b). Compared with most previous research, our study reveals the comparison the two approaches and how they can be combined for rating prediction.

First, our findings suggest that compared with sentiments, the topic memberships of latent dimensions generated from the review text have better performance in rating prediction. The top membership could be considered as a helpful tool to predict customer overall satisfaction. Furthermore, the heterogeneity of the predictive power exists across different dimensions. Several dimensions have good performance in rating prediction while several dimensions don't show good enough performance. The new features we extracted from review text can be collaborated with customer sentiment to achieve better performance in rating prediction both holistically and individually. It extends the literature by combining the topic membership with customer sentiment instead of only adopting one feature from the review text.

Second, several models have been proposed by researchers to achieve higher accuracy in rating prediction (Cheng et al., 2018; Xing et al., 2019). Compared with these approaches, our approach has more flexibility and interpretability, especially when considering points of intervention in the customer facing areas of the business. The use of topic modelling approach allows us to extract latent dimensions from the textual data without pre-labelling the data, which saves human efforts of training and adjusting the model. The adoption of unsupervised machine learning approach (LDA) saves the human effort to train the model. Besides, the adoption of Shapley value could clearly show how each dimension extracted from review text contributes to predicting the overall ratings, which provides more interpretability of how each latent dimensions and customer sentiment affect the customer overall satisfaction.

5.3 Practical implications

Customer ratings are direct measurements of customer overall satisfaction while the textual content represents customer perception towards experience showing customer satisfaction indirectly (predicting overall). Findings from this study could provide several managerial insights for restaurant owners and managers. First, our findings provide restaurants with identification of latent dimensions from a large amount of customer reviews, which demonstrates customers' various aspects of perception towards food and delivery service. Both

praise and complaint from customers could help restaurant owners to develop operation strategies.

Customer feedback is a vital information source to understand customers' opinions, which has been proved to have significant influence on customer behaviour and sales (Li et al., 2019; Z. Zhao et al., 2019a, 2019b). The 15 topics extracted from review text shows the multidimensionality of customers' evaluation. For instance, Topic #13 and Topic #8 illustrates customers' complaint about long delivery time and order issues respectively while Topic #2 represents customers praise towards food quality.

Furthermore, the heterogeneity which exists in the contributions of each topic membership to customer overall satisfaction, may help restaurants to prioritize the most influencing factors. By identifying the most important positive and negative dimensions that influence the rating scores, restaurant managers could explicit dissatisfaction and enhance their weakness accordingly or develop marketing strategies by highlighting their strength. For instance, our findings show that except for polarity, Topic #13 mainly illustrating the long delivery time contributes most to predicting rating score. And the long delivery time is the most important factor negatively affecting the rating score minimizing the reviews lead to higher rating score. Following an identification of the contribution of each service quality factor, an owner may also identify priorities and monitor improvement by incorporating other covariates such as service time and other inputs such as quality of raw materials etc.

6 Conclusions, limitations, and future research

Finding the underlying reasons for rating scores using contextual information is critical for businesses to develop strategies to discover why customers have different levels of satisfaction. To discover the value of unstructured text within customer reviews to explain actual customer ratings, we evaluate and compare how two approaches (*sentiment analysis* and *topic models*) can be applied to understand customer satisfaction. This study demonstrates that incorporating document-level covariates, such as topic membership, can greatly contribute to the understanding of the sentiment of customer feedback, such as the ones found in customer reviews, and predict the review score in a much better way than the actual tone of the review text, even in cases where access to sentiment vocabulary may be limited. While a large body of literature has demonstrated that review text is primarily consistent with the rating—and therefore, the sentiment of the review is reflected in the star rating of this particular review, from a business owner's point of view—latent dimensions discovered from these review texts are a useful instrument to be incorporated in the business practice. They can identify areas of improvement and competency that the business can expand the most.

Nevertheless, there are several limitations to our research. For the robustness check in our first experiment, we only classify reviews into positive and negative classes, not considering the neutral class, which has been commonly studied in online review literature. Furthermore, even though we have examined two granularity levels of sentiment, we only employ a single dictionary. The choice of dictionaries may have a different influence on prediction accuracy. Therefore, these aspects could be improved in future work.

Future work should focus on several directions. First, reviews classified as neutral could be considered, together with the positive and negative ones, as three distinct classes for classifying customer rating scores to different satisfaction levels. Second, apart from lexicon-based methods, other unsupervised machine-learning techniques could be adopted to detect customers' polarity scores. Also, the subjectivity and emotions contained in the review text

could be considered in future work. Third, customer reviews from other platforms could be included in the future to discover the level of heterogeneity across different platforms and different service domains.

Declarations

Conflict of interests The authors disclose no financial or conflict of interests.

References

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications, 6*(1). https://doi.org/10.14569/IJACSA.2015.060121

Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management, 54*, 102132. https://doi.org/10.1016/j.ijinfomgt.2020.102132

Bagheri, A., Saraee, M., & De Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems, 52*, 201–213.

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications, 127*, 256–271. https://doi.org/10.1016/j.eswa.2019.03.001

Batra, S., & Bawa, S. (2010). Using lsi and its variants in text classification. *Advanced techniques in computing sciences and software engineering* (pp. 313–316). Springer.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226*, 107134.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022 https://dl.acm.org/doi/10.5555/944919.944937

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

Brintrup, A. (2021). *AI in the supply chain: a classification framework and critical analysis of current state*. In Oxford handbook of supply chain management: OUP, USA. https://doi.org/10.1093/oxfordhb/9780190066727.013.24

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

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, (pp. 288–296).

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2021). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research, 131*, 815–825.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

Cheng, Z., Ding, Y., Zhu, L., & 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).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.
Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*(1), 61–84. https://doi.org/10.3166/DN.17.1.61-84
Dey, A., Jenamani, M., & Thakkar, J. J. (2018). Senti-N-Gram: An n-gram lexicon for sentiment analysis. *Expert Systems with Applications, 103*, 92–105.
Do, H. H., Prasad, P., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications, 118*, 272–299.
Elshakankery, K., & Ahmed, M. F. (2019). HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis. *Egyptian Informatics Journal, 20*(3), 163–171.
Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model. *Applied Sciences, 11*(23), 11091.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels—An empirical analysis. *Tourism Management, 61*, 43–54.
Ghasemaghaei, M., Eslami, S. P., Deal, K., & Hassanein, K. (2018). Reviews' length and sentiment as correlates of online reviews' ratings. *Internet Research, 28*(3), 544–563. https://doi.org/10.1108/IntR-12-2016-0394
Ghiassi, M., & Lee, S. (2018). A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications, 106*, 197–216.
Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498–1512.
Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C., & Voyiatzis, I. (2021). XGBoost and deep neural network comparison: The case of teams' performance. In *International Conference on Intelligent Tutoring Systems*, (pp. 343–349).
Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the Twenty-Fourth Annual Conference of Cognitive Science Society*, (pp. 381–386).
Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management, 59*, 467–483.
Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1), 177–196.
Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, (pp. 755–760).
Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems, 57*, 42–53.
Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management, 72*, 417–426.
Khanam, Z., Alwasel, B., Sirafi, H., & Rashid, M. (2021). Fake news detection using machine learning approaches. In *IOP Conference Series: Materials Science and Engineering*, (pp. 012040).
Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications, 152*, 113401.
Koltcov, S., Koltsova, O., & Nikolenko, S. (2014). Latent dirichlet allocation: stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science*, (pp. 161–165). https://doi.org/10.1145/2615569.2615680.
Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications, 116*, 472–486.
Kumar, A., Gopal, R. D., Shankar, R., & Tan, K. H. (2022). Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. *Decision Support Systems, 155*, 113728.
Kwon, H.-J., Ban, H.-J., Jun, J.-K., & Kim, H.-S. (2021). Topic modeling and sentiment analysis of online review for airlines. *Information, 12*(2), 78.
Lai, X., Wang, F., & Wang, X. (2021). Asymmetric relationship between customer sentiment and online hotel ratings: The moderating effects of review characteristics. *International Journal of Contemporary Hospitality Management, 33*(6), 2137–2156. https://doi.org/10.1108/IJCHM-07-2020-0708
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.

Li, X., Wu, C., & Mai, F. (2019). The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information & Management, 56*(2), 172–184.
Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of natural language processing*. Oxfordshire (Vol. 2, pp. 627–666).
Liu, B. (2012). Sentiment analysis and opinion mining. In *Synthesis lectures on human language technologies* (Vol. 5, pp. 1–167). https://doi.org/10.2200/S00416ED1V01Y201204HLT016)
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, (pp. 4768–4777).
Mai, L., & Le, B. (2021). Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations Research, 300*(2), 493–513. https://doi.org/10.1007/s10479-020-03534-7
Marshan, A., Kansouzidou, G., & Ioannou, A. (2020). Sentiment analysis to support marketing decision making process: A hybrid model. In *Proceedings of the future technologies conference*, (pp. 614–626).
McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on recommender systems*, (pp. 165–172).
Minka, T. P., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, (pp. 352–359).
Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models interpretable*. Lulu.
Pang, B., Lee, L., & 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*, https://doi.org/10.3115/1118693.1118704.
Parasuraman, A., Zeithaml, V. A., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40.
Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 569–577).
Qiu, J., Liu, C., Li, Y., & Lin, Z. (2018). Leveraging sentiment analysis at the aspects level to predict ratings of reviews. *Information Sciences, 451*, 295–309. https://doi.org/10.1016/j.ins.2018.04.009
Qu, L., Ifrim, G., & Weikum, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, (pp. 913–921).
Quan, C., & Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences, 272*, 16–28.
Rao, V. C. S., Radhika, P., Polala, N., & Kiran, S. (2021). Logistic regression versus XGBoost: Machine learning for counterfeit news detection. In *2021 second international conference on smart technologies in computing, electrical and electronics (ICSTCEE)*, (pp. 1–6).
Rinker, T. (2020). qdap: Bridging the gap between qualitative data and quantitative analysis. *R Package Version, 2*(4), 3.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.
See-To, E. W., & Ngai, E. W. (2018). Customer reviews for demand distribution and sales nowcasting: A big data approach. *Annals of Operations Research, 270*(1), 415–431.
Seo, S., Huang, J., Yang, H., & Liu, Y. (2017). Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the eleventh ACM conference on recommender systems*, (pp. 297–305).
Sharma, S. S., & Dutta, G. (2021). SentiDraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination. *Information Processing & Management, 58*(1), 102412.
Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE international conference on data science and advanced analytics (DSAA)*, (pp. 165–174). https://doi.org/10.1109/DSAA.2017.61.
Tan, Y., Zhang, M., Liu, Y., & Ma, S. (2016). Rating-boosted latent topics: Understanding users and items with ratings and reviews. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, (pp. 2640–2646).
Tirunillai, S., & 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. https://doi.org/10.1509/jmr.12.0106
Verma, S., & Yadav, N. (2021). Past, present, and future of electronic word of mouth (EWOM). *Journal of Interactive Marketing, 53*, 111–128.
Wang, Y., Zhong, K., & Liu, Q. (2022). Let criticism take precedence: Effect of side order on consumer attitudes toward a two-sided online review. *Journal of Business Research, 140*, 403–419.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, (pp. 29–39).

Wu, J., Li, Y., & Ma, Y. (2021). Comparison of XGBoost and the neural network model on the class-balanced datasets. In *2021 IEEE 3rd international conference on frontiers technology of information and computer (ICFTIC)*, (pp. 457–461).

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51–65.

Xing, S., Wang, Q., Zhao, X., & Li, T. (2019). A hierarchical attention model for rating prediction by leveraging user and product reviews. *Neurocomputing, 332*, 417–427.

Xu, X. (2020). Examining an asymmetric effect between online customer reviews emphasis and overall satisfaction determinants. *Journal of Business Research, 106*, 196–210.

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review, 53*(6), 4335–4385.

Yan, Z., Wang, J., Dong, Q., Zhu, L., Lin, W., & Jiang, X. (2022). XGBoost algorithm and logistic regression to predict the postoperative 5-year outcome in patients with glioma. *Annals of Translational Medicine, 10*(16), 860–860.

Yeo, S. F., Tan, C. L., Kumar, A., Tan, K. H., & Wong, J. K. (2022). Investigating the impact of AI-powered technologies on Instagrammers' purchase decisions in digitalization era—A study of the fashion and apparel industry. *Technological Forecasting and Social Change, 177*, 121551.

Zelikovitz, S., & Hirsh, H. (2001). Using LSI for text classification in the presence of background text. In *Proceedings of the tenth international conference on Information and knowledge management*, (pp. 113–118).

Zhang, H., Shen, F., Liu, W., He, X., Luan, H., & Chua, T.-S. (2016). Discrete collaborative filtering. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*, (pp. 325–334).

Zhang, C., Tian, Y.-X., & Fan, L.-W. (2020). Improving the Bass model's predictive power through online reviews, search traffic and macroeconomic data. *Annals of Operations Research, 295*(2), 881–922.

Zhang, W., & Wang, J. (2016). Integrating topic and latent factors for scalable personalized review-based rating prediction. *IEEE Transactions on Knowledge and Data Engineering, 28*(11), 3013–3027.

Zhao, Y., Xu, X., & Wang, M. (2019a). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management, 76*, 111–121.

Zhao, Z., Wang, J., Sun, H., Liu, Y., Fan, Z., & Xuan, F. (2019b). What factors influence online product sales? Online reviews, review system curation, online promotional marketing and seller guarantees analysis. *IEEE Access, 8*, 3920–3931.