EXPLAINING THE STARS: ASPECT-BASED SENTIMENT ANALYSIS OF ONLINE CUSTOMER REVIEWS

Research paper

Binder, Markus, University of Regensburg, Regensburg, Germany, Markus1.Binder@ur.de

Heinrich, Bernd, University of Regensburg, Regensburg, Germany, Bernd.Heinrich@ur.de

Klier, Mathias, University of Ulm, Ulm, Germany, Mathias.Klier@uni-ulm.de

Obermeier, Andreas, University of Regensburg, Regensburg, Andreas.Obermeier@ur.de

Schiller, Alexander, University of Regensburg, Regensburg, Germany, Alexander.Schiller@ur.de

Abstract

The importance of online customer reviews for the success of products and services has been recognized in both research and practice. Therefore, the ability to explain and interpret customer assessments expressed by the assigned overall star ratings is an important and interesting research field. Existing approaches for explaining the overall star ratings, however, often do not address methodical issues associated with these ratings (e.g., ordinal scale). Moreover, they often ignore the review texts which contain valuable information on the customers' assessments of different aspects of the rated items (e.g., price or quality). To contribute to both research gaps, we propose a generalized ordered probit model using aspect-based sentiments as independent variables to explain the overall star ratings of online customer reviews. For measuring the explanatory power of our model, we suggest a likelihood-based pseudo R-squared measure. By evaluating our approach using a large real-world dataset of restaurant reviews we show, that, in contrast to other regression models, the generalized ordered probit model can address the methodical issues associated with the star ratings. Moreover, the evaluation shows that the results of the proposed model are easy to interpret and valuable for analysing customer assessments.

Keywords: Online customer reviews, Explanatory model, Aspect-based sentiment analysis, Generalized ordered probit model.

1 Introduction

In recent years, the number of internet users has increased from 1,024 million in 2005 up to 3,578 million in 2017 (ITU, 2017). This increase has considerably contributed to the rise of popular platforms such as Amazon (Linden et al., 2003) or TripAdvisor (Filieri et al., 2015) which, inter alia, provide access to online customer reviews (O'Mahony and Smyth, 2010). Online customer reviews can be an important instrument to reduce information asymmetries about offered products and services (Hu et al., 2008). They contain rich information about customers' assessments and opinions in form of user generated content (Ye et al., 2011) and typically consist of an overall star rating (e.g., 1 to 5 stars) and a textual part (Mudambi et al., 2014). The overall star ratings summarize the customers' general impressions of the rated items. The textual parts comprise further details on the customers' assessments, often towards different aspects of the rated items (e.g., service quality in a restaurant review), to justify and explain the associated overall ratings (Zhu et al., 2011). Indeed, literature already provides some approaches to analyse these textual assessments in terms of aspect-based sentiments (Schouten and Frasincar, 2016).

Online customer reviews may affect the economic success of products and services considerably (e.g., Chevalier and Mayzlin, 2006; Clemons et al., 2006; Minnema et al., 2016; Phillips et al., 2017; Ye et al., 2009; Ye et al., 2011; Zhu and Zhang, 2010). Research has shown that besides high overall star ratings, positive feedback contained in the textual parts reviews yields, amongst others, higher sales volumes (Archak et al., 2007, 2011; Ghose and Ipeirotis, 2011). Even though the analysis of structural data, such as star ratings or metadata on the items, is predominantly focussed by existing literature, the textual parts of reviews have been shown to comprise very valuable information (Ganu et al., 2013). In that line, some predictive models have been proposed (e.g., Goldberg and Zhu, 2006; Li et al., 2011; Pang and Lee, 2005; Qu et al., 2010) which aim to predict the star ratings based on review texts. However, these models mostly rely on latent variables which are hard to interpret as they do not necessarily represent the thematic aspects focussed by the users when reviewing the item. Indeed, explaining and interpreting the overall star ratings based on such predictive models is not aimed at or possible. To make the rich information contained in the review texts accessible, an explanatory model is needed, which uses easy to interpret independent variables like aspect-based sentiments. Such an explanatory model enables the identification of causal relationships between the independent variables (i.e., aspect-based sentiments) and the dependent variable (i.e., the associated overall star rating) (Sainani, 2014).

Aspect-based sentiment analysis accounts for the review texts including the users' assessments of different aspects of the rated items in a methodically well-founded way (Jo and Oh, 2011; Schouten and Frasincar, 2016; Zhu et al., 2011). In that line, we use aspect-based sentiments contained in the review texts and propose an approach to explain and interpret the users' overall star ratings. We focus on the following research question:

> *How can aspect-based sentiments contained in the textual parts of online customer reviews be used to explain and interpret the associated overall star ratings?*

To answer this question, we aim at an explanatory model (cf. Shmueli, 2010; Shmueli and Koppius, 2011) to explain the associated overall star ratings based on easy to interpret aspect-based sentiments. We argue that the principles and the knowledge base of regression theory are adequate and valuable, providing well-founded methods to analyse and explain the associated overall star ratings of online customer reviews. In general, results of a regression analysis are easy to interpret as they allow to understand how the dependent variable (i.e., the overall star rating) changes on average, when the independent variables (i.e., the aspect-based sentiments) are varied (Myers, 1990). However, focusing on the given problem definition, the application of a regression analysis faces different methodical issues associated with the star ratings. Amongst others, these methodical issues arise from their ordinal scale (e.g., 1 to 5 stars as integer). To address such methodical issues and in contrast to existing approaches, we base our approach on a generalized ordered probit regression model. From a scientific point of view, the proposed approach aims to uncover the underlying reasoning of the overall star ratings as it uses interpretable aspect-based sentiments given in the review texts avoiding any latent variables. For practitioners, our model enables companies to gain a data-driven competitive advantage by being able to analyse the reasoning behind customer ratings and customer assessments. Such an explanation for the users' overall star ratings allows for customer orientation based on the evidence and importance of different item aspects which are relevant for customer (dis)satisfaction. For example, businesses could focus their efforts on actions to improve on those aspects which influence users' (dis)satisfaction most. Thus, the presented approach provides a way to explain overall star ratings based on the review texts not yet targeted by existing approaches, resolves the associated methodical issues, and is relevant to research and practice.

The remainder of the paper is structured as follows: In the next section, we discuss both the related literature and the research gap. In Section 3, we step-by-step develop our model for explaining star ratings using aspect-based sentiments. In Section 4, we demonstrate and evaluate our approach using a large dataset of restaurant reviews. Section 5 depicts implications of our approach for theory and practice. Finally, we conclude, reflect on limitations and provide an outlook on further research.

2 Related Work and Research Gap

In this section, we analyse existing research which aims at *explaining* overall star ratings of online customer reviews using regression models. Thereby, we also consider works using structural and textual (item) data different from aspect-based sentiments as they might be interesting from a methodological point of view. Existing contributions with a sole predictive (or descriptive) perspective such as Pang and Lee (2005), Qu et al. (2010), Li et al. (2011), Zhou et al. (2014), Tang et al. (2015), Monett and Stolte (2016), Sharma et al. (2016) or Qiu et al. (2018), do not aim to explain or interpret the (overall) star ratings and are thus out of scope for our research. These works are not considered in the following.

In accordance with the guidelines of standard approaches to prepare the related work (e.g., Levy and Ellis, 2006; Webster and Watson, 2002), we searched the databases ScienceDirect, Google Scholar, ACM Digital Library, EBSCO Host, IEEE Xplore, and the AIS Library for the following search term and without posing a restriction on the time period: *("regression" and rating*) or ("regression" and review*) or ("regression" and "recommender")*. Additionally, we performed a forward and backward search starting from highly relevant papers. The papers found were manually screened based on title, abstract, keywords and summary. The 51 papers remaining after this first screening were analysed in detail and 11 of them were identified as relevant for our work.

	Consideration of aspect-based sentiments	**Addressing methodical issues (e.g., the ratings' ordinal scale)**	**Evaluation of the explanatory power of the model**
Approaches considering structural (item) data			

Guo et al. (2016);
Liu et al. (2017);
Radojevic et al. (2017);
Ye et al. (2014)
	n/a	n/a	✓

Yang et al. (2018)	n/a	✓	n/a
Approaches considering textual (item) data			
Fu et al. (2013); Linshi (2014)	n/a	n/a	n/a
Debortoli et al. (2016); Xiang et al. (2015)	n/a	n/a	✓
Ganu et al. (2009); Ganu et al. (2013)	✓	n/a	n/a

Table 1. Existing approaches for explaining the overall star ratings of online customer reviews

Table 1 provides an overview of the identified papers. They contribute to the problem of modelling the overall star ratings of online customer reviews using regression models with different sets of independent variables (i.e., structural (item) data or textual (item) data). The respective approaches are grouped depending on the characteristic of these independent variables (highlighted by different shades and subheadings). The first column of Table 1 states whether aspect-based sentiments are considered. The second column indicates whether the proposed regression models address methodical issues relevant in the context of explaining overall star ratings. For example, it is necessary to consider the fact that the dependent variable (i.e., the overall star rating) is ordinally scaled (i.e., discrete and ordered) (Debortoli et al., 2016). The third column states whether the explanatory power of the regression model is evaluated using a well-founded quality measure (e.g., the explained variance).

Guo et al. (2016), Liu et al. (2017), Radojevic et al. (2017) and Ye et al. (2014) use regression models with structural data as independent variables to model the overall star ratings of reviews and evaluate the explanatory power of their models by calculating (adjusted) R-squared values. Radojevic et al. (2017) propose a linear multi-level regression model for overall star ratings, using structural data regarding the items (e.g., *price* or *free internet*) and the users (e.g., regarding nationality or travel experience) as independent variables. Guo et al. (2016), Liu et al. (2017) and Ye et al. (2014) use sub-ratings explicitly given by the users (e.g., room experience and service on a 5-point Likert scale). Thereby, Guo et al. (2016) and Liu et al. (2017) analyse the relationships between explicitly given sub-ratings as independent variables and the overall rating as dependent variable in the hotel domain. Ye et al. (2014) investigate the relationship between price as independent variable and given sub-ratings for service quality or value as dependent variable. All four works – Guo et al. (2016), Liu et al. (2017), Radojevic et al. (2017) and Ye et al. (2014) – provide first insights in the underlying reasons for customer assessments in online customer reviews. However, in none of these works aspect-based sentiments in the review texts are used. Instead, Guo et al. (2016), Liu et al. (2017) and Ye et al. (2014) rely on explicitly given sub-ratings. In reality such explicitly given multi-ratings represent an exceptional case limiting these approaches to some extent. Moreover, all four works use common linear regression models which do not address the methodical issues that arise when explaining the overall star ratings. In particular, the ordinal scale of the star ratings is not considered. Neglecting such methodical issues may lead to significant misspecifications and thus invalid results. Yang et al. (2018) are the only ones to account for the methodical issue of ordinally scaled overall ratings. They introduce an ordinal regression model to infer the overall star ratings from structural location-based data of items (i.e., hotels). Their aim is to explain a hotel’s guest assessments (given by the average rating of the hotel) based on information about the hotel’s location (e.g., accessibility to points of interest or the location’s surrounding environment). The approach relies on structural data regarding the location and the authors do not aim at using review texts or aspect-based sentiments. Additionally, they do not assess the explanatory power of their model, which is a challenging problem, as there are no standard quality measures for the presented ordinal regression model. Summing up, the approaches using structural data are hampered in their applicability (assumption that sub-ratings are given) and/or by the missing consideration of the methodical issues associated with the star ratings (e.g., the ordinal scale) and/or the respective evaluation of the explanatory power of the model. Additionally, they do not take advantage of the review texts or aspect-based sentiments.

Indeed, there also exist approaches using independent variables derived from textual (item) data to explain the star ratings of online customer reviews. Fu et al. (2013) and Linshi (2014) propose linear regression models to explain the associated star ratings. Thereby, Fu et al. (2013) employ word counts based on the review texts as independent variables. Linshi (2014) use document vectors from a codeword Latent Dirichlet Allocation (LDA) which is able to distinguish different topics based on the connotation (good vs. bad) of the co-occurring words (e.g., good food vs. bad food). However, in both works, the authors use linear regression models which do not account for the methodical issues associated with the overall star ratings like their ordinal scale. Additionally, they do not further investigate the explanatory power of the proposed regression models. Debortoli et al. (2016) and Xiang et al. (2015) indeed analyse the explanatory power of their regression models based on the review texts. Debortoli et al. (2016) – similar to Linshi (2014) – use document vectors from a LDA based on the review texts as explanatory variables. They provide a multinomial logistic regression model for explaining the associated overall star ratings. To assess the explanatory power of their model, the deviance explained is stated. Xiang et al. (2015) propose a linear regression model based on the factor loadings from a factor analysis of the review texts. The explanatory power is assessed in terms of the adjusted R-squared measure. The methodical issues, however, are not addressed in both approaches, as the ordinal scale of the star ratings is neglected. In addition, document vectors from a LDA (Debortoli et al., 2016) or factor loadings (Xiang et al., 2015), respectively, do not necessarily account for (different) sentiments. For example, different sentiments may be contained in one single topic or factor (e.g., one topic or factor concurrently containing statements for good and bad food) or one sentiment may be distributed over different topics or factors. This weakens the interpretability resp. validity of the results. To conclude, the approaches for explaining the overall star ratings of reviews discussed in this paragraph do not address the methodical issues associated with the overall star ratings. In particular, the ordinal scale of the star ratings is neglected. Moreover, they do not account for aspect-based sentiments.

Ganu et al. (2009) and Ganu et al. (2013) show that aspect-based sentiments contained in review texts can be used to improve recommender systems. Both papers generally focus on *predicting* a user’s star rating for a restaurant based on his or her previous ratings for other restaurants and the ratings of all other users. However, in minor parts of the papers (i.e., Section 3.3 of Ganu et al. (2009) and Section 3.2

of Ganu et al. (2013)) regression models for inferring the associated overall star ratings using aspect-based sentiments are discussed. These regressions are based on sentence types, represented as (aspect, sentiment)-pairs assigned to every sentence. To construct the sentence types, each sentence of the review texts is classified according to one aspect it most probably refers to (e.g., *food*, *service* or *miscellaneous*). Additionally, a sentiment label (e.g., *positive*, *neutral* or *negative*) is assigned to each sentence. On this basis, multivariate regression models for the associated overall star ratings are proposed using sentence type fractions in the review texts as independent variables. More precisely, a sentence type fraction is calculated as the percentage of sentences of that type contained in the review text. Ganu et al. (2009) use a linear and Ganu et al. (2013) a quadratic regression model. Both, however, focus on using aspect-based sentiments to improve recommender systems but do not aim at explaining and interpreting the associated overall star ratings. Therefore, they do not further investigate the explanatory power of the proposed regression models (e.g., in terms of coefficients of determination). Additionally, the allocation of sentiment labels is equivalent to a classification instead of a more fine-grained representation of the sentiments as numerical values. Finally, the authors apply common regression models, which do not address the methodical issues associated to the star ratings (e.g., the ordinal scale).

To conclude, there are very interesting contributions regarding modelling the overall star ratings of online customer reviews which can serve as a basis for further research. To uncover the causal relationships between aspect-based sentiments contained in review texts and the associated overall star ratings, an explanatory model is needed. However, existing literature lacks an explanatory model using aspect-based sentiments to explain the associated overall star ratings which addresses the occurring methodical issues (e.g., ordinal scale of the star ratings). Furthermore, the explanatory power of (different sets of) aspect-based sentiments has not been investigated yet. Due to the methodical issues arising, amongst others from the ordinal scale of the star ratings, this is particularly challenging.

3 A Model to Explain Star Ratings

To address this research gap, we propose an explanatory model for overall star ratings with respect to aspect-based sentiments, which addresses the methodical issues associated with the star ratings. We first introduce the basic idea of our approach. Then, we outline a generalized ordered probit model for the analysis of star ratings. Finally, we propose a likelihood-based pseudo R-squared measure for assessing the explanatory power of aspect-based sentiments in this context.

3.1 Basic idea of our approach

Our aim is to explain the overall star ratings of textual reviews based on the associated aspect-based sentiments. To do that, first, an adequate regression model addressing the methodical issues for modelling star ratings has to be established. These issues result in particular from both the ordinal scale of star ratings and the characteristics of aspect-based sentiments. Then, the explanatory power of different aspect-based sentiments can be assessed using this model.

Our approach is based on the ordered probit model (McKelvey and Zavoina, 1975). To adequately represent star ratings, we follow a two-step approach. First, an underlying model for continuous preferences instead of discrete star ratings is established (Greene and Hensher, 2010). Then, a non-linear transformation of the underlying preferences onto the rating scale is used. More precisely, the ratings are modelled by dividing the underlying continuous preference variable into intervals of different size.

To elaborate why this two-step approach is proposed, we discuss different methodical issues for modelling star ratings. Thereby, we compare the ordered probit model to a linear regression model because the latter is commonly used in literature (cf. Section 2). First and crucially, an ordered probit model accounts for the ordinal scale of the star ratings, whereas a linear regression model does not and thus might lead to significant misspecifications. To achieve an accurate representation, an explanatory model has to reflect *uneven distances within the (ordinal) rating scale.* For instance, on a scale from 1 to 5 a rating of 4 might, on average, be much closer to a rating of 5 with respect to the underlying preference than to a rating of 3 (Greene and Hensher, 2010). A linear regression model is not able to cope with this issue, whereas the ordered probit model accounts for uneven distances within the rating scale by assigning preference intervals of different sizes to the ratings. Further, a model for star ratings must cope with a *non-normal distribution of the rating errors* (due to the star ratings being discrete) and with *heteroscedasticity of the ratings* (due to the bounded scale of the star ratings). In contrast to a linear regression model, our proposed approach addresses these issues by estimating unbounded continuous preferences in a first step. Finally, *varying impacts of the aspect-based sentiments* over the rating scale might occur. For instance, in the context of a restaurant review, a poor service (e.g., due to an unfriendly waiter) may easily lead to assigning the lowest rating, but a pleasant service alone will in general not be sufficient to assign the highest rating. This can be taken into account by generalizing the ordered probit model to allow varying coefficients of the aspect-based sentiments.

3.2 Generalized ordered probit model to analyse aspect-based sentiments

We consider a set of $M \in \mathbb{N}$ textual reviews. Each review is associated with a star rating r on a discrete scale from 1 to a maximal rating of $K \in \mathbb{N}$. This is the common review structure observed for popular platforms such as Amazon or TripAdvisor (with K=5 or K=10 for most platforms). For each review, we take into account $A \in \mathbb{N}$ different item aspects relevant regarding the associated star rating. To give an example, in a restaurant review possible item aspects might be food quality or service quality. For instance, in the review “The food was great” a strongly positive sentiment towards the aspect food quality is expressed. More generally, we analyse the sentiment $s_a \in \mathbb{R}$ towards each item aspect $a \in 1, \ldots, A$. In this way, a numerical value is assigned to the sentiment s_a. Overall, for review i (with $i \in \{1, \ldots, M\}$) this results in aspect-based sentiments $s_1^i, \ldots, s_A^i \in \mathbb{R}$ and an associated star rating $r^i \in \{1, \ldots, K\}$.

In our two-step approach, first, preferences $R_*^i \in \mathbb{R}$ are modelled using the aspect-based sentiments $s_1^i, ., s_A^i$. Later, the preferences are transformed into ratings in a non-linear way. According to the classical ordered probit model, the underlying preferences are given by

$$R_*^i = \beta_1 s_1^i + \ldots + \beta_A s_A^i + \epsilon, \tag{1}$$

where $\beta_1, \ldots, \beta_A$ denote the parameters with respect to the aspect-based sentiments $s_1^i, \ldots, s_A^i$ and $\epsilon \sim N(0,1)$ denotes the random error term of the underlying linear preference model reflecting the ambiguity contained in textual reviews (Mudambi et al., 2014). To account for the uncertainty stemming from the error term, we also introduce a discrete random variable $R^i \in \{1, \ldots, K\}$ to estimate the actual rating r^i in the i-th review. In the underlying linear preference model, the intercept term can be omitted since flexible threshold terms $\theta_1 < \ldots < \theta_{K-1} \in \mathbb{R}$ are used to transform the preferences into ratings, i.e., $R^i = 1$ for $R_*^i \leq \theta_1, R^i = 2$ for $\theta_1 < R_*^i \leq \theta_2, \ldots, R^i = K$ for $R_*^i > \theta_{K-1}$.

The parameters $\beta_1, \ldots, \beta_A$ and the thresholds $\theta_1, \ldots, \theta_{K-1}$ have to be estimated according to the classical ordered probit model. To give an example, consider a set of restaurant reviews on a rating scale from 1 to 5 addressing only the sentiments towards food quality and service. Then, an exemplarily resulting model might be given by the preference model $R_*^i = 1.0 \cdot s_{food}^i + 0.5 \cdot s_{service}^i + \epsilon$ (i.e., $\beta_1 = 1.0$ and $\beta_2 = 0.5$) and the non-linear transformation $R^i = 1$ if $R_*^i \leq -2.5 (= \theta_1), R^i = 2$ if $-2.5 < R_*^i \leq -0.8\ (= \theta_2), \ldots, R^i = 5$ if $R_*^i > 3.2\ (= \theta_4)$ onto the rating scale.

Those parameters are fitted by maximizing the log-likelihood of the model. According to the preference model in Equation (1) and the transformation onto the rating scale as introduced above, it is given by

$$\begin{aligned} &\log L(\beta_1, ., \beta_A, \theta_1, ., \theta_{K-1}) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{K} Z_{ij} \log[\Phi(\theta_j - \beta_1 s_1^i - \ldots - \beta_A s_A^i) - \Phi(\theta_{j-1} - \beta_1 s_1^i - \ldots - \beta_A s_A^i)], \end{aligned} \tag{2}$$

where $Z_{ij} = 1$ if $r^i = j$, $Z_{ij} = 0$ otherwise, $\theta_0 := -\infty$, $\theta_K := +\infty$ and Φ denotes the cumulative distribution function of the standard normal distribution. That is, the likelihood of a rating j in the i-th review is given by $P(R^i = j) = P(R^i \leq j) - P(R^i \leq j-1)$ in the model, which means, by the difference in the cumulative probability to the next lowest rating.

In Equation (2), $P\big(R^i \leq j\;\big|s_1^i, \ldots, s_A^i\big) = P\big(R_*^i \leq \theta_j \big| s_1^i, \ldots, s_A^i\big) = \Phi(\theta_j - \beta_1 s_1^i - \ldots - \beta_A s_A^i)$ is assumed. In other words, the parameters $\beta_1, \ldots, \beta_A$ are independent of the rating value j ('Parallel Lines Assumption'). However, for example, a positive price-sentiment towards an item may have different impacts: Its impact might be stronger when the rating is at least mediocre on a rating scale from 1 to 5 (i.e., on $P\big(R^i \geq 3\big) = 1 - P(R^i \leq 2)$), whereas it might be lower when the associated rating is very good (i.e., on $P\big(R^i = 5\big) = 1 - P(R^i \leq 4)$). More generally, the Parallel Lines Assumption has to be tested for each aspect-based sentiment $s_a \in \{s_1, \ldots, s_A\}$. If it does not hold for s_a, a relaxed version

$$P\big(R^i \leq j\;\big|s_1^i, \ldots, s_A^i\big) = P\big(R_*^i \leq \theta_j \big| s_1^i, \ldots, s_A^i\big) = \Phi(\theta_j - \ldots - \beta_a^j s_a^i - \ldots) \tag{3}$$

with different coefficients β_a^j has to be used.

To test the Parallel Lines Assumption for sentiment s_a, the Bayesian Information Criterion (BIC) can be used (Schwarz, 1978). The assumption holds if $\log(M)\,(K-2) > 2 \cdot \big(\log L(\widehat{\beta_{G_a}}, \widehat{\theta_{G_a}}) - \log L(\hat{\beta}, \hat{\theta})\big)$, where $\hat{\beta}$, $\hat{\theta}$ and $\widehat{\beta_{G_a}}$, $\widehat{\theta_{G_a}}$ are the maximum likelihood estimates for the classical version and the relaxed version G_a for sentiment s_a, respectively. Since the BIC takes into account the sample size M, it copes with the problem that for large samples, the model with more degrees of freedom often "falsely" gives distinctly higher likelihoods due to overfitting. As the sample sizes for the analysis of textual reviews are typically very high, the BIC is generally a well-suited measure in this context. Overall, our ordered probit model is generalized to varying coefficients for every aspect-based sentiment violating the Parallel Lines Assumption.

3.3 Measure to assess the explanatory power for the proposed model

In the following, we propose a measure to assess the explanatory power of different aspect-based sentiments for our generalized ordered probit model. To do so, we assess the explained variability by different aspect-based sentiments in the underlying linear preference model. Thereby, the variability explained by the underlying preference model (i.e., its R-squared value) can be identified with its likelihood. More precisely, in this case the R-squared value can be evaluated by

$$\mathcal{R}^2 = 1 - \left[\frac{L_{Null-Model}}{L_{Preference-Model}}\right]^{2/M}, \tag{4}$$

where $L_{Preference-Model}$ denotes the likelihood of the fitted preference model (Maddala, 1983). Similarly, $L_{Null-Model}$ denotes the likelihood of a preference model restricted to $\beta = 0$. That is, the null model yields a constant preference regardless of the aspect-based sentiments. For the proposed generalized model, this identification of the R-squared value matches the explained variability in each preference model for $R^i \leq j$ and thus provides a well-founded overall estimate of the variability explained by the underlying generalized preference model.

However, our generalized ordered probit model for star ratings includes an additional variance, since the exact preferences underlying the assigned star ratings are unknown. That is, the likelihood of the underlying preference model is not directly accessible. To cope with this issue and to take into account the additional variance of the preference distribution, we propose to rescale the measure to have a maximum value of 1 for the generalized ordered probit model. In Nagelkerke (1991), this rescaling of the R-squared measure was already proposed for models that are fitted by maximum likelihood estimation in general, but in our generalized ordered probit model it is especially suited. Since our approach indeed includes underlying linear preference models, the measure inherits the precise foundation in Equation (4) when applied to our generalized ordered probit model. Overall, in our context the proposed measure is given by

$$\mathcal{R}^2_{Nagelkerke} = \frac{1 - \left[\frac{L_{Null-Model}}{L_{Gen.Ord.Prob.-Model}}\right]^{2/M}}{1 - L_{Null-Model}{}^{2/M}}. \tag{5}$$

This Nagelkerke pseudo R-squared measures how likely our generalized ordered probit model based on aspect-based sentiments is, compared to a null-model that does not factor in the aspect-based sentiments at all (i.e., restricting all coefficients of the aspect-based sentiments to zero in Equation (2)). In that way, it assesses the variability explained by the underlying preference model (Veall and Zimmermann, 1992). Having established an R-squared-type measure for the proposed model, we are able to evaluate the proposed model on different subsets of reviews and thereby gain valuable insights on the impact of different aspect-based sentiments.

4 Evaluation

In this section we evaluate our proposed model on a large dataset of restaurant reviews. First, we discuss the reasons for selecting the dataset and describe its preparation. Then, we methodically evaluate our approach in comparison to alternative models on our real-world dataset. Finally, we present the results of our proposed model for selected sentiment aspects.

4.1 Case selection and preparation of the dataset

To evaluate our approach, we use a large real-world dataset of reviews for restaurants in New York City from 2010-2017 provided by an established web portal for online customer reviews regarding local businesses, especially restaurants. Overall, the dataset consists of 2.4 million textual restaurant reviews and their associated star ratings. The characteristics of the dataset are summarized in Table 2. Thereby, the density of available reviews (calculated as the number of reviews divided by the product of the numbers of users and items) and the skewness of the rating distribution (‘J-shaped’) are in line with previous literature (e.g., Askalidis et al., 2017; Debortoli et al., 2016; Huang et al., 2004). Since these characteristics are typical for online customer reviews and since the dataset is large enough to analyse different sentiment aspects (each with a sufficient number of reviews), we selected this real-world dataset to apply and evaluate our model.

First, aspect-based sentiments have to be extracted from the reviews in the dataset. This step is necessary to apply the proposed generalized ordered probit model. However, it is not part of our contribution of the paper at hand (thus, it is described as dataset preparation). To extract sentiments from text, well-established methods exist (Agarwal et al., 2015; Liu, 2012; Taboada et al., 2011). Thereby, supervised learning approaches and dictionary-based approaches can be distinguished (Liu, 2012). Since supervised learning approaches require manual labelling of a large number of reviews, we decided to use a dictionary-based approach as in (Taboada et al., 2011). It is, however, important to note that generally supervised learning approaches may also be used to determine the inputs for our proposed model. For our evaluation, we applied separate sentiment dictionaries for different aspects in the restaurant context. This allowed us to account for varying sentiment orientations depending on the referred aspect. For example, the word “low” has a positive sentiment when referring to the price, whereas its sentiment orientation is negative for other aspects (e.g., “low food quality”).

For our evaluation and without any loss of generality, we considered the aspects price, service, food quality, ambience, food quantity and miscellaneous. These aspects are broadly consistent with literature (e.g., Kiritchenko et al., 2014), but generally, additional aspects or separations (such as food quality vs. food quantity) may also be included as inputs for our model. To account for these different aspects in our analysis, we determined the referred aspect for every word expressing a sentiment in the reviews. Therefore, we used a list of index words for each considered aspect. Then, we applied the Stanford NLP Dependency Parser (Schuster and Manning, 2016), as in Kiritchenko et al. (2014) and Agarwal et al. (2015), to match sentiment words appearing in the review texts with the corresponding index words. For example, in the sentence “The *waitress* was *friendly*.” The sentiment word *friendly* is matched with the index word *waitress*, which refers to the aspect service. Moreover, we aggregated the mean sentiment for each aspect accounting for intensified, weakened and negated contexts (Taboada et al., 2011). The implementation was done in Python. Finally, to avoid unstable results by an explanatory model, multicollinearity between the extracted aspect-based sentiments was tested to be sufficiently low. This is underlined by a variance inflation factor (VIF) of 1.12 in Table 2 (i.e., max. 1-1/1.12=11% of an aspect-based sentiment can be explained by sentiments towards other aspects) (Mansfield and Helms, 1982; O’brien, 2007).

Characteristics of the dataset	
# of users / restaurants	583’815 / 18’507
# of textual reviews and ratings	2'396'643
# of users with high review count (>50)	5’146
# of restaurants with high review count (>100)	6’197
Considered aspect-based sentiments	price, service, food quality, ambience, food quantity, and miscellaneous
Multicollinearity between the aspect-based sentiments measured by the VIF	1.12

Table 2. Characteristics of the dataset

4.2 Methodical evaluation of our approach

Having prepared the dataset, the sentiments $s_1^i, \ldots, s_6^i$ (towards price, service, food quality, ambience, food quantity, miscellaneous) and the associated rating r^i ($\in \{1, \ldots, 5\}$) are given for each review ($\in \{1, \ldots, 2'396'643\}$). Based on this real-world dataset, we evaluate the ability of different approaches to address the methodical issues discussed in Section 3.1. More precisely, we compare the ordered probit model and its proposed generalized version to a linear regression model because the latter is commonly used in literature to model and explain star ratings (cf. Section 2).

For the classical ordered probit model we get, according to Equation (1), the preference model

$$R_*^i = \beta_1 s_1^i + \ldots + \beta_6 s_6^i + \epsilon,$$

with $\epsilon \sim N(0,1)$ and the strictly non-linear transformation onto the rating scale

$$R^i = 1 \text{ for } R_*^i \leq \theta_1, R^i = 2 \text{ for } \theta_1 < R_*^i \leq \theta_2, \ldots, R^i = 5 \text{ for } R_*^i > \theta_4.$$

The proposed generalized ordered probit model can formally be written as

$$R^i \leq j \text{ if } \beta_1^j s_1^i + \ \beta_2^j s_2^i + \ldots + \beta_6^j s_6^i + \epsilon \leq \theta_j \text{ for } j = 1,2,3,4$$

with $R^i \in \{1, \ldots, 5\}$, $\epsilon \sim N(0,1)$ and different coefficients $\beta_a^1, \beta_a^2, \beta_a^3, \beta_a^4$ instead of one fixed coefficient β_a for the aspect-based sentiments $s_a \in \{s_1, \ldots, s_6\}$ that violate the Parallel Lines Assumption.

Using a linear regression the ratings are modelled as

$$R^i = \theta_0 + \beta_1 s_1^i + \ldots + \beta_6 s_6^i + \epsilon$$

with an intercept θ_0 and an error term $\epsilon \sim N(0, \sigma^2)$.

As already discussed in Section 3.1., these models differ in their ability to address the methodical issues for modelling star ratings with respect to aspect-based sentiments. In the following, we evaluate these three models regarding the four methodical issues discussed in Section 3.1:

First, we examined whether *uneven distances within the (ordinal) rating scale* exist on the dataset. Therefore, we analysed the overall sentiment (defined as $s_1 + s_2 + \ldots + s_6$) of each review in the dataset. More precisely, we determined the average value of the overall sentiment $s_1 + s_2 + \ldots + s_6$ over all reviews grouped by the assigned star rating. Having determined these values, the distance between two star ratings can be identified with the difference in the average overall sentiment expressed in the corresponding reviews. Thereby, for example, the increase in this value from a 4-star to a 5-star review was detected to be less than half compared to all other adjacent star ratings. More precisely, the standardized differences (to have an average value of 1) in the overall sentiments amount to 1.1 (1 to 2 stars), 1.3 (2 to 3 stars), 1.1 (3 to 4 stars) and only 0.5 (4 to 5 stars). In that line, indeed uneven distances can be detected on our dataset. Thus, the assumption of even distances within the (ordinal) rating scale made by the linear regression model is not met. In contrast, the classical and the generalized ordered probit model can cope with uneven distances by assigning preference intervals of different sizes to the ratings.

Second, we determined whether a *non-normal distribution of the rating errors* occurs on our dataset. Therefore, we performed a Kolmogorov-Smirnov test (Massey, 1951) of the normality assumption for the linear regression model, which failed on the dataset. The test gave a vanishing probability that the cumulative distribution of the error term stems from a normal distribution ($p < 10^{-16}$). Hence, the assumption of normally distributed errors made by the linear regression model is not valid. In contrast, the (generalized) ordered probit models do not assume a specific distribution of the rating errors.

Third, we examined whether *heteroscedasticity of the ratings* is an issue in our dataset. This can be detected by comparing the linear model to a relaxed version with a scalable error variance $\beta_v \widehat{R^\iota}$ instead of a fixed error variance σ^2, where β_v denotes the additional variance parameter and $\widehat{R^\iota}$ the estimated rating. Adding the variance parameter β_v leads to an improvement of $3'620$ in the BIC which reveals the presence of heteroscedasticity. Hence, the assumption of homoscedasticity of the rating in the linear regression model is not met. In contrast, the classical and the generalized ordered probit model are not hampered by such an assumption and thus can handle the occurring heteroscedasticity of the ratings.

Finally, we tested whether *varying impacts of the aspect-based sentiments* can be detected in our dataset. To uncover possible varying impacts, we compared (similarly to above) the differences within the rating scale, but separately for different aspect-based sentiments. Thereby, for instance, the standardized differences in the service sentiment amount to 1.5 (1 to 2 stars), 1.1 (2 to 3 stars), 0.9 (3 to 4 stars) and 0.5 (4 to 5 stars). This indicates that the aspect-based sentiments indeed have significantly varying impacts since, for instance, the service sentiment differs over-proportionally between 1- and 2-star ratings (1.5 vs. distance 1.1 overall, as detected by analysing overall uneven distances above). That is, a model assuming a constant coefficient for each aspect-based sentiment, such as the linear regression model, is strongly limited in its validity. To verify that our proposed model captures these different impacts, we also compared the classical ordered probit model to a generalized version by the respective BIC values. Thereby, also significant varying impacts over the rating scale were detected by a difference in the BIC value of 2’686. Since Raftery (1995) defined differences bigger than 10 already as ‘very strong evidence’ for the model with the lower BIC value, the proposed generalized version is more valid.

Overall, the methodical evaluation above shows that indeed all of the methodical issues discussed in Section 3.1 occur on our dataset. Our proposed model is able to address these issues, whereas the classical ordered probit model does not account for varying impacts of the aspect-based sentiments and the linear regression model does not resolve any of the discussed issues. Table 3 summarizes the results.

	Accounts for uneven distances within the (ordinal) rating scale	**Allows for non-normal distribution of the rating errors**	**Accounts for heteroscedasticity of the ratings**	**Accounts for varying impacts of the aspect-based sentiments**	**BIC (relative to the Generalized Ordered Probit Model)**
Ordered Probit Model	✔	✔	✔	n/a (constancy assumed)	2’686
Generalized Ordered Probit Model	✔	✔	✔	✔	-
Linear Regression Model	n/a (even distances assumed)	n/a (normal distribution assumed)	n/a (homoscedasticity assumed)	n/a (constancy assumed)	39’029
Empirical evidence for methodical	Related standardized differences are significantly uneven (from 0.5 to 1.3)	Kolmogorov-Smirnov test rejects normal distribution	Additional variance parameter in linear model leads to a more	Impacts of certain sentiments (e.g., service sentiment)	

issues in our dataset		assumption ($p < 10^{-16}$)	valid model (i.e. higher BIC)	differ significantly between 1- and 2-star ratings	

Table 3. Comparison of different regression models on the dataset

To further evaluate the considered regression models, we also compared the relative quality of these models. Therefore, the values of the BIC relative to the generalized ordered probit model are also stated in Table 3. As mentioned in Section 3.2, the BIC accounts for the sample size and thus is suited for large datasets such as our dataset of restaurant reviews. Thereby, the values of the BIC indicate that the proposed generalized ordered probit model is methodically much better suited to explain the star ratings on our dataset than a classical ordered probit and especially a linear regression model.

4.3 Results for selected aspect-based sentiments

In this section, we present and discuss the results for selected aspect-based sentiments based on our dataset. Since our main contribution addresses methodical issues on explaining the star ratings and due to length restrictions, we limited ourselves to the (three most frequently referred) aspect-based sentiments for price, service and food quality.

At first, we built our model based on the subset of reviews expressing sentiments towards all three of these aspects (and do not address any of the other extracted aspects). Price, service and food quality sentiment have different impacts over the rating scale (i.e., violate the Parallel Lines Assumption), which underlines the relevance of our proposed generalized ordered probit model. This model is given by

$$R^i \leq j \text{ if } \beta_{price}^j s_{price}^i + \ \beta_{service}^j s_{service}^i + \beta_{food}^j s_{food}^i + \epsilon \leq \theta_j \text{ for } j = 1,2,3,4.$$

Coefficients	$j = 1$	$j = 2$	$j = 3$	$j = 4$
θ_j (threshold)	-0.89	-0.07	0.78	1.59
β_{price}^j (price sentiment)	0.11	0.16	0.17	0.11
$\beta_{service}^j$ (service sentiment)	0.28	0.29	0.28	0.22
β_{food}^j (food quality sentiment)	0.29	0.38	0.43	0.34

Table 4. Coefficients in the generalized ordered probit model (based on reviews that address the price, service and food quality sentiment)

The coefficients of the model are provided in Table 4. The Nagelkerke pseudo R-squared (cf. Equation (5)) is 44%. All coefficients were highly statistically significant ($p < 10^{-7}$). Overall, we noticed that the price sentiment has a lower impact than the service or food quality sentiment. One reason might be that the price level of the restaurant is often known prior to the visit while the service quality and the food quality are experienced during the stay. We found that the food quality sentiment indeed has the strongest impact on the overall preferences (in terms of the assigned star ratings). Surprisingly though, the sentiment towards service has a similarly strong impact on the preferences for parts of the rating scale. In particular, in the lowest rating category, the service sentiment has an (almost) equally strong impact as the food quality sentiment (0.28 vs 0.29 in Table 4). This indicates that for poorly rated restaurants a lacking service quality is equally bad as a low food quality. Notably, we would not have been able to detect that characteristic using a linear regression model or the classical ordered probit model. In the latter models, fixed coefficients are estimated, whereas in our proposed model the ratio of the coefficients for food quality sentiment and service sentiment vary significantly (for example, 0.29/0.28 = 1.04 for $j = 1$ vs. 0.34/0.22 = 1.55 for $j = 4$).

To further evaluate the benefits of the proposed model, we compared it to the classical ordered probit model, which does not allow for varying impacts. Thereby, the Nagelkerke pseudo R-squared is 42% for the ordered probit model. That is, on a general level the explanatory power is nearly the same as for our model. However, to gain more detailed insights in the differences in validity, we analysed how well the two models are able to explain the ratings for different parts of the rating scale. Thereby, the results might be biased by the skewed rating distribution (‘J-shaped’). To eliminate this bias, we randomly sampled an equal number of 1,000 reviews for each rating category and built the models on this sample. Thereby, the proposed model had significantly higher explanatory power measured by the Nagelkerke pseudo R-squared for high and low ratings: In detail, 71% vs. 63% for rating 1, 38% vs. 27% for rating 2, 24% vs. 29% for rating 3, 45% vs. 44% for rating 4 and 69% vs. 63% for rating 5. Except of the average rating of 3, the proposed model explains the ratings more accurately for all rating categories. Overall, this indicates that the proposed generalized ordered probit model outperforms the classical ordered probit model by additionally addressing varying impacts of the aspect-based sentiments (cf. Table 3).

Besides the reviews addressing all three aspect-based sentiments, there are reviews which address only one or two of them. In that line different subsets of reviews can be identified (based on the addressed aspect-based sentiments). To examine and compare how well our proposed model is able to explain the overall star ratings on these subsets, we evaluated the respective Nagelkerke pseudo R-squared values given in Table 5. All coefficients in all models were highly statistically significant ($p < 10^{-7}$).

Aspects addressed in the reviews	Nagelkerke pseudo R-squared by our proposed model
Price / Service / Food	20% / 49% / 32%
Price & Service / Price & Food / Service & Food	48% / 35% / 43%
Price & Service & Food	44%

Table 5. Nagelkerke pseudo R-squared based on reviews addressing different sentiment aspects

By establishing the Nagelkerke pseudo R-squared as an estimator for the explained variability in the generalized ordered probit model, we have an evaluation measure to compare the explanatory power of aspect-based sentiments on different subsets of reviews. That is, we have a replacement for the R-squared measure in a methodically sound model for star ratings and thus get a more valid comparison of the different subsets, compared to using a linear regression model.

Thereby, we found that for reviews addressing only one of these three sentiments, the service sentiment even explains the most variability in the star ratings (49% vs. 32% for the food quality sentiment and 20% for the price sentiment). This indicates that reviews only addressing the service often express a definite sentiment towards that aspect (e.g., complaints about unfriendly service). For reviews that address the food quality sentiment we found that the ones containing additional sentiment aspects explain the star ratings better (44% vs. 32% with food quality sentiment alone). This indicates that one-dimensional reviews towards the food quality often do not discuss additional aspects that influence their overall preference towards the restaurant. For reviews addressing the service sentiment we detected a different effect. The ones containing additional sentiments tended to explain the star ratings slightly worse in comparison (44% vs. 49% with service sentiment alone). According to the first observation, this might be due to the fact that reviews addressing mainly the service often express a definite sentiment towards that aspect which strongly affects the associated overall rating, while reviews addressing multiple aspects might only mention the service for the sake of completeness. Overall, using our approach we detected significantly varying impacts of the aspect-based sentiments both within the rating scale and (in terms of the explained variability) based on the combination of aspects addressed in the reviews.

5 Implications for Theory and Practice

In contrast to existing approaches for explaining the star ratings of online customer reviews, our approach takes advantage of the valuable information contained in aspect-based sentiments which are measured in the review texts. Furthermore, it addresses the methodical issues which emerge during the explanation of overall star ratings, particularly due to their ordinal scale. In that way, by applying our approach the impact of aspect-based sentiments on the associated overall star ratings can be explained and interpreted in a methodically well-founded way. Proposing a generalized ordered probit model and allowing the consideration of different sets of aspect-based sentiments, our approach can provide a detailed level of analysis. For example, as indicated by the evaluation of the model on a large real-world dataset of restaurant reviews, valuable insights about varying impacts of aspect-based sentiments on the overall star ratings can be discovered. Finally, having established a quality criterion for the proposed model in form of a R-squared type measure, our approach is able to compare the strength of the relationships between different sets of aspect-based sentiments and the associated overall star ratings.

From a methodical point of view, the results presented in Section 4 indicate that our proposed model is methodically better suited than a linear regression, which is commonly used in state-of-the-art approaches, as well as a classical ordered probit model to explain the overall star ratings of online customer reviews. This is especially supported by the fact that our dataset is large enough to be representative and exhibits characteristics typical for online customer reviews (cf. Section 4.1). Moreover, the methodical issues occurring when explaining overall star ratings (i.e., uneven distances within the (ordinal) rating scale, non-normal distribution of the rating errors, heteroscedasticity of the ratings and varying impacts of the aspect-based sentiments; cf. Table 3) are addressed by our proposed approach. Overall, the relative BIC values in Table 3 are all significantly large, which indicates, that compared to the alternative models our approach is able to explain some additional substantial part of the variation of star ratings.

Our proposed model can help practitioners to gain a data-driven competitive advantage by using the aspects and the associated sentiments to analyse why specific customer ratings were assigned to their company or a competitor. Based on these insights the company can innovate its business model and further develop customer-centric solutions adding business value. This competitive advantage can be achieved in diverse areas of application such as decision support, quality management or marketing. Using the detailed information about the aspects influencing the customer assessment, businesses can focus their efforts on actions in a more effective and target-oriented way. For example, the use of financial, infrastructural and human resources can be improved with respect to customer demands. In quality management, our approach allows to identify reasons for a possible drop in customer satisfaction and thus enables suitable countermeasures. Using our approach, marketing analysts can study the reasons for customer (dis)satisfaction on a detailed level. This allows them to ensure customer orientation by considering client needs and meeting their major priorities.

6 Conclusion, Limitations and Future Work

Explaining the underlying reasoning for the overall star ratings of online customer reviews is an important issue in both research and practice. In this paper, we present an approach to explain and interpret the overall star ratings of online customer reviews using aspect-based sentiments contained in review texts. The proposed approach contributes to existing research by allowing for a detailed and interpretable understanding of the customer assessment of products and services. We propose a generalized ordered probit model and a Nagelkerke pseudo R-squared measure to explain the overall star ratings using aspect-based sentiments. Existing approaches lack such an explanatory regression model addressing the methodical issues associated with the star ratings and assessing the explanatory power for the model. A formal definition of the approach was provided, and it was evaluated on a large real-world dataset of restaurant reviews. The evaluation was conducted in two steps. In a first step, we methodically evaluated our approach by comparing the proposed model to alternative regression models on our dataset. Therein, we showed, that our approach is able to address the methodical issues occurring when explaining overall star ratings of online customer reviews. In a second step, we presented the results of our proposed model for selected sentiment aspects. Thereby, our approach yields interpretable results and detailed relationships between aspect-based sentiments and the overall star ratings have been uncovered.

Nevertheless, our work also has limitations which may constitute the starting point for future research. In this paper we focused on evaluating the explanatory power of given sets of aspect-based sentiments. Future research could explore the usage of sentiments towards automatically extracted, but interpretable aspects. Furthermore, the approach was applied to a large real-world dataset from the restaurant domain. The fact that the dataset that our dataset is large and exhibits typical characteristics for online customer reviews suggests that the results will apply in other domains in a similar way. Nevertheless, future research could evaluate it on further datasets from other domains. Finally, further evaluations and methodical extensions (e.g., considering additional structural data such as given sub-ratings or item data) could also provide interesting insights regarding the explanation of star ratings of online customer reviews.

References

Agarwal, B., N. Mittal, P. Bansal and S. Garg (2015). “Sentiment analysis using common-sense and context information” *Computational intelligence and neuroscience* 2015, 715–730.

Archak, N., A. Ghose and P. G. Ipeirotis (2007). “Show me the money!: deriving the pricing power of product features by mining consumer reviews”. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 56–65.

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

Askalidis, G., S. J. Kim and E. C. Malthouse (2017). “Understanding and overcoming biases in online review systems” *Decision Support Systems* 97, 23–30.

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

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

Debortoli, S., O. Müller, I. A. Junglas and J. Vom Brocke (2016). “Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial” *Communications of the Association for Information Systems* 39, 110–135.

Filieri, R., S. Alguezaui and F. McLeay (2015). “Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth” *Tourism Management* 51, 174–185.

Fu, B., J. Lin, L. Li, C. Faloutsos, J. Hong and N. Sadeh (2013). “Why people hate your app: Making sense of user feedback in a mobile app store”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1276–1284.

Ganu, G., N. Elhadad and A. Marian (2009). “Beyond the stars: improving rating predictions using review text content”. In: *Twelfth International Workshop on the Web and Databases (WebDB 2009)*, pp. 1–6.

Ganu, G., Y. Kakodkar and A. Marian (2013). “Improving the quality of predictions using textual information in online user reviews” *Information Systems* 38 (1), 1–15.

Ghose, A. and P. G. Ipeirotis (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.

Goldberg, A. B. and X. Zhu (2006). “Seeing stars when there aren’t many stars: graph-based semi-supervised learning for sentiment categorization”. In: *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pp. 45–52.

Greene, W. H. and D. A. Hensher (2010). *Modeling Ordered Choices. A Primer:* Cambridge University Press.

Guo, Y., S. Barnes and Q. Jia (2016). “Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation” 59, 467–483.

Hu, N., L. Liu and J. J. Zhang (2008). “Do online reviews affect product sales? The role of reviewer characteristics and temporal effects” *Information Technology and management* 9 (3), 201–214.

Huang, Z., H. Chen and D. Zeng (2004). “Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering” *ACM Transactions on Information Systems (TOIS)* 22 (1), 116–142.

ITU, I. T.U. (2017). *World telecommunication/ICT indicators database:* International Communication Union Geneva, Switzerland.

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

Kiritchenko, S., X. Zhu, C. Cherry and S. Mohammad (2014). “NRC-Canada-2014: Detecting aspects and sentiment in customer reviews”. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442.

Li, F., N. Liu, H. Jin, K. Zhao, Q. Yang and X. Zhu (2011). “Incorporating reviewer and product information for review rating prediction”. In: *IJCAI*, pp. 1820–1825.

Linden, G., B. Smith and J. York (2003). “Amazon. com recommendations: Item-to-item collaborative filtering” *IEEE Internet computing* (1), 76–80.

Linshi, J. (2014). “Personalizing Yelp star ratings: A semantic topic modeling approach” *Yale University*.

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

Liu, Y., T. Teichert, M. Rossi, H. Li and F. Hu (2017). “Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews” *Tourism Management* 59, 554–563.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics:* Cambridge University Press.

Mansfield, E. R. and B. P. Helms (1982). “Detecting Multicollinearity” *The American Statistician* 36 (3a), 158–160.

Massey, F. J. (1951). “The Kolmogorov-Smirnov Test for Goodness of Fit” *Journal of the American Statistical Association* 46 (253), 68–78.

McKelvey, R. D. and W. Zavoina (1975). “A statistical model for the analysis of ordinal level dependent variables” *The Journal of Mathematical Sociology* 4 (1), 103–120.

Minnema, A., T. H. A. Bijmolt, S. Gensler and T. Wiesel (2016). “To keep or not to keep: effects of online customer reviews on product returns” *Journal of retailing* 92 (3), 253–267.

Monett, D. and H. Stolte (2016). “Predicting Star Ratings based on Annotated Reviews of Mobile Apps”. In: *FedCSIS*, pp. 421–428.

Mudambi, S. M., D. Schuff and Z. Zhang (2014). “Why aren’t the stars aligned? An analysis of online review content and star ratings”. In: *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pp. 3139–3147.

Myers, R. H. (1990). *Classical and modern regression with applications:* Duxbury press Belmont, CA.

Nagelkerke, N. J. D. (1991). “A Note on a General Definition of the Coefficient of Determination” *Biometrika* 78 (3), 691–692.

O’brien, R. M. (2007). “A Caution Regarding Rules of Thumb for Variance Inflation Factors” *Quality & Quantity* 41 (5), 673–690.

O’Mahony, M. P. and B. Smyth (2010). “Using readability tests to predict helpful product reviews”. In: *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pp. 164–167.

Pang, B. and L. Lee (2005). “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales”. In: *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124.

Phillips, P., S. Barnes, K. Zigan and R. Schegg (2017). “Understanding the impact of online reviews on hotel performance: an empirical analysis” *Journal of Travel Research* 56 (2), 235–249.

Qiu, J., C. Liu, Y. Li and Z. Lin (2018). “Leveraging sentiment analysis at the aspects level to predict ratings of reviews” *Information Sciences* 451, 295–309.

Qu, L., G. Ifrim and G. Weikum (2010). “The bag-of-opinions method for review rating prediction from sparse text patterns”. In: *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 913–921.

Radojevic, T., N. Stanisic and N. Stanic (2017). “Inside the rating scores: A multilevel analysis of the factors influencing customer satisfaction in the hotel industry” *Cornell Hospitality Quarterly* 58 (2), 134–164.

Raftery, A. E. (1995). “Bayesian model selection in social research” *Sociological methodology*, 111–163.

Sainani, K. L. (2014). “Explanatory versus predictive modeling” *PM&R* 6 (9), 841–844.

Schouten, K. and F. Frasincar (2016). “Survey on aspect-level sentiment analysis” *IEEE Transactions on Knowledge and Data Engineering* 28 (3), 813–830.

Schuster, S. and C. D. Manning (2016). “Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks”. In: *LREC*, pp. 23–28.

Schwarz, G. (1978). “Estimating the dimension of a model” *The Annals of Statistics* 6 (2), 461–464.

Sharma, R. D., S. Tripathi, S. K. Sahu, S. Mittal and A. Anand (2016). “Predicting online doctor ratings from user reviews using convolutional neural networks” *International Journal of Machine Learning and Computing* 6 (2), 149.

Shmueli, G. (2010). “To explain or to predict?” *Statistical science* 25 (3), 289–310.

Shmueli, G. and O. R. Koppius (2011). “Predictive analytics in information systems research” *MIS Quarterly*, 553–572.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede (2011). “Lexicon-Based Methods for Sentiment Analysis” *Computational Linguistics* 37 (2), 267–307.

Tang, D., B. Qin, T. Liu and Y. Yang (2015). “User Modeling with Neural Network for Review Rating Prediction”. In: *IJCAI*, pp. 1340–1346.

Veall, M. and K. Zimmermann (1992). “Pseudo-R2's in the ordinal probit model” *The Journal of Mathematical Sociology* 16 (4), 333–342.

Xiang, Z., Z. Schwartz, J. H. Gerdes Jr and M. Uysal (2015). “What can big data and text analytics tell us about hotel guest experience and satisfaction?” *International Journal of Hospitality Management* 44, 120–130.

Yang, Y., Z. Mao and J. Tang (2018). “Understanding guest satisfaction with urban hotel location” *Journal of Travel Research* 57 (2), 243–259.

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

Ye, Q., R. Law, B. Gu and W. Chen (2011). “The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings” *Computers in Human behavior* 27 (2), 634–639.

Ye, Q., H. Li, Z. Wang and R. Law (2014). “The influence of hotel price on perceived service quality and value in e-tourism: An empirical investigation based on online traveler reviews” *Journal of Hospitality & Tourism Research* 38 (1), 23–39.

Zhou, L., S. Ye, P. L. Pearce and M.-Y. Wu (2014). “Refreshing hotel satisfaction studies by reconfiguring customer review data” *International Journal of Hospitality Management* 38, 1–10.

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

Zhu, J., H. Wang, M. Zhu, B. K. Tsou and M. Ma (2011). “Aspect-based opinion polling from customer reviews” *IEEE Transactions on Affective Computing* 2 (1), 37–49.