THE POWER OF TEXT: INVESTIGATING THE EXPLANATORY POWER OF TEXTUAL ONLINE CONSUMER REVIEWS FOR STAR RATINGS

Completed Research Paper

Markus Binder, University of Regensburg, Regensburg, Germany, markus1.binder@ur.de

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

Marcus Hopf, University of Regensburg, Regensburg, Germany, marcus.hopf@ur.de

Michael Szubartowicz, University of Regensburg, Regensburg, Germany, michael.szubartowicz@ur.de

Abstract

Online consumer reviews are important for the success of consumer products. To analyse star ratings, existing research has leveraged the expectancy-disconfirmation theory (EDT), but by deriving consumer expectations from prior star ratings rather than aspect-based sentiments in prior textual online reviews. The latter is promising since it can allow actionable insights for different product aspects such as service of restaurants. To operationalize an aspect-level research model, we leverage large language models in this work. Our results support, amongst others, that expectations, perceived performance and disconfirmation, each derived from textual reviews, have a unique explanatory power for star ratings with perceived performance having the highest one for all three considered product domains. This study contributes to a more thorough understanding of the explanatory power of textual online consumer reviews for star ratings and gives actionable insights for businesses to efficiently improve the average star rating as key performance indicator.

Keywords: Online Consumer Reviews, Star Ratings, Expectancy-Disconfirmation Model, Aspect-based Sentiment Analysis, eWOM.

1 Introduction

There are hundreds of millions of online consumer reviews available on web portals such as Google, Amazon or Yelp (Rubin, 2024). They comprise rich information (Jabr et al., 2018) and typically consist of a star rating (e.g., 1 to 5 stars) representing the overall consumer assessment and a textual part. The importance of online consumer reviews for the success of products and services is widely recognized (Gutt et al., 2019; Jabr & Rahman, 2022). That is, research has shown that online consumer reviews considerably impact sales (Jabr & Rahman, 2022). In that line, online reviews are crucial information sources for other consumers. More precisely, 75% of consumers “always” or “regularly” read online reviews when researching local businesses according to a recent survey (Paget, 2024). Here, star ratings are a crucial component of online reviews, as only 29% of consumers would consider using a business with an average rating below 3.5 stars (Paget, 2024). Thus, star ratings are important performance indicators for businesses offering consumer products (Gutt et al., 2019).

The textual parts of online consumer reviews comprise further consumer assessments towards different aspects of the rated product such as service for restaurants or battery for laptops (Siering & Janze, 2019). These are referred to as aspect-based sentiments and can be extracted from the texts to determine which product aspects are especially important for consumers (Chatterjee et al., 2021; Jabr et al., 2018). They have shown to impact sales beyond the overall star rating and are thus of particular interest for companies (Sahoo et al., 2018). The increasing relevance of text analytics is further accelerated by the research on large language models such as BERT or GPT, which led to breakthrough performances on a variety of information extraction tasks from text (Devlin et al., 2019; J. Yang et al., 2023). In that line, analysing textual online consumer reviews has been recognized as important area in research on electronic word-of-mouth (eWOM), which studies statements of individuals available on the internet about objects of their interest (Donthu et al., 2021; Jabr et al., 2020).

Online consumer reviews and especially the assigned star ratings can yield important insights on consumer satisfaction (Chatterjee et al., 2021). Here, the expectancy-disconfirmation theory (EDT) (Oliver, 2014) suggests that *expectations*, *perceived performance* and *disconfirmation* (i.e., the difference between perceived performance and expectations) are the basis for post-purchase evaluations. We denote these three different components of the EDT as EDT dimensions in this work. Focusing on star ratings, a major source for expectations are prior online consumer reviews and especially the textual parts (Jha & Shah, 2019). While existing eWOM research has investigated expectations based on prior star ratings, analysing a modelling framework in which an individual's expectations are derived from prior textual reviews on a sentiment-aspect-level is a promising research direction (Chatterjee et al., 2021; Nam et al., 2020). That is, such analyses can allow detailed insights for different product aspects without the need for time-consuming and costly consumer surveys (Chatterjee et al., 2021). Hereto, it is important to investigate whether a research model loses explanatory power by leaving out a specific EDT dimension. This is referred to as unique explanatory power of the EDT dimension for star ratings in this work. If all three EDT dimensions indeed have a unique explanatory power for star ratings, this would support that no EDT dimension derived from textual reviews is fully mediated. This contributes to an important theoretical discussion in the literature on the EDT about different possibilities of mediation effects between the EDT dimensions (Oliver, 2014). Moreover, such analyses can give valuable insights for businesses to efficiently allocate resources to improve specific product aspects or to establish suited marketing strategies for managing consumer expectations. To that end, we formulate four hypotheses for the unique explanatory power of the three EDT dimensions, each derived from textual reviews, for star ratings. To extract aspect-based sentiments from a large number of review texts, we apply the encoder language model BERT (Devlin et al., 2019). Given this set of extracted aspects of different dimensions as independent variables, star ratings as dependent variable are analysed by using an ordered probit model (Binder et al., 2019). To gain broader and more robust insights, we consider three different types of goods with variability along the established distinction between experience goods and search goods in the literature (Jin et al., 2023; Mudambi & Schuff, 2010). For instance, restaurants are generally considered as experience goods since they require direct product experience for a proper assessment, while for search good domains such as laptops most of the relevant product aspects can be assessed prior to the purchase. By means of a research model on a sentiment-aspect level for online consumer reviews, we thus focus on the following research questions in this study:

> RQ1: Do the EDT dimensions *expectations*, *perceived performance* and *disconfirmation*, each derived from textual reviews, have a unique explanatory power for star ratings in online consumer reviews?
>
> RQ2: Which of these EDT dimensions has the highest or lowest unique explanatory power for star ratings?
>
> RQ3: Does the unique explanatory power of each of the EDT dimensions for star ratings differ between experience goods and search goods?

Analysing three datasets of online consumer reviews, our results support that all three EDT dimensions have a unique explanatory power on star ratings with *perceived performance* consistently having the highest one in our research model. Further, we find that *expectations* derived from prior textual reviews have a relatively high unique explanatory power for movies.

To the best of our knowledge, we are the first to investigate the EDT based on expectations derived from prior textual reviews on a sentiment-aspect-level in online reviews. Moreover, we are the first to investigate the unique explanatory power of the three EDT dimensions *expectations*, *perceived performance* and *disconfirmation* on star ratings in a research model on a sentiment-aspect-level. Here, deriving the EDT dimensions from textual reviews is considered a promising research area (Nam et al., 2020). By our aspect-level research model, we also contribute to the identified research gap of broadly one-dimensional textual analyses in existing research (Baden et al., 2022; Liu et al., 2024).

2 Background

In this section, we first introduce relevant theory and then discuss the related work.

2.1 Theory

Existing research suggests that post-purchase consumer evaluations are influenced by *expectations* and *perceived performance* (Ho et al., 2017; Hult et al., 2022). To that end, the expectancy-disconfirmation theory (EDT) is increasingly popular (Oliver, 2014; J. Zhang et al., 2022). It considers the dimensions *expectation*, *perceived performance* and *disconfirmation*. Here, disconfirmation generally describes the discrepancy between perceived performance and expectation. If consumers experience a large deviation between *perceived performance* and *expectations*, existing literature on the EDT suggests that these large positive or negative disconfirmations are especially relevant for explaining post-purchase evaluations (Ho et al., 2017; Oliver, 2014). That is, some consumers tend to assign extreme star ratings then (Ho et al., 2017). This might stem from consumers' reactions, driven by a need to influence public opinion by either expressing strongly divergent views or by being particularly motivated to post extreme reviews. (Han & Anderson, 2020). In contrast, small disconfirmations are considered to have a quite low impact on star ratings and are sometimes even analysed to be within a "zone of indifference" in the literature (Oliver, 2014). Overall, existing research suggests that each of the EDT dimensions *expectations*, *perceived performance* and *disconfirmation* individually is related to post-purchase evaluations (Oliver, 2014; J. Zhang et al., 2022). Here, mediation effects between the EDT dimensions are theoretically discussed in the literature as *disconfirmation* is per definition connected to *expectations* and *perceived performance* (Oliver, 2014). As an extension to the EDT framework, it has also been suggested that *expectations*, *perceived performance* and *disconfirmation* are built separately for different aspects of a product (Oliver, 2014). This could allow to gain detailed insights for different product aspects within the EDT framework. The EDT has already become the predominant approach for explaining consumer satisfaction in related fields such as public services (J. Zhang et al., 2022). For online consumer reviews, it is also commonly used for analysing dynamics in star ratings of a product (Heinrich et al., 2022). These insights can be used by practitioners, for instance, to develop appropriate marketing strategies (Jha & Shah, 2021).

2.2 Related Work

To discuss related research investigating star ratings of online consumer reviews, we embed our study into the field of eWOM. Regarding the framework of eWOM literature by Jabr et al. (2020), our research can be classified as evaluation of eWOM focusing on the investigation of the semantic and lexical content of online consumer reviews. That is, we investigate the explanatory power of textual reviews for consumers' star ratings decisions. Thus, existing studies focusing on the preliminary step of relevant factors for consumers to decide whether to post a review or subsequent analyses of the impact of online consumer reviews on sales are out of scope for our research (Qahri-Saremi & Montazemi, 2022; Ren & Nickerson, 2019).

There are several recent studies which analyse star ratings in online consumer reviews by means of aspect-based sentiments (Binder et al., 2019; Chatterjee et al., 2021; Jabr et al., 2018; Khan et al., 2022; Kolomoyets & Dickinger, 2023). These studies generally focus on approaches to extract aspect-based sentiments from review texts and on their impact on the corresponding star rating. That is, they focus on the relationship between the perceived performance expressed in the review text and the star rating within an online consumer review, but do not take into account expectations from prior reviews. Thus, these studies also do not investigate the unique explanatory power of these different EDT dimensions.

Indeed, there is also existing research on the impact of prior online consumer reviews on subsequent reviews (Guo & Zhou, 2016; Ma et al., 2013). Here, for instance, Guo and Zhou (2016) analyse the impact of the amount of prior reviews and their variance in star ratings on subsequent star ratings. Yet, they focus on prior star ratings only and not on the impact of prior textual reviews. Ma et al. (2013) analyse how review length and time interval since the last review moderate their impact on subsequent reviews, but they also do not take into account the semantical content of the review texts.

Finally, there are studies which leverage the EDT to investigate star ratings in online consumer reviews (Heinrich et al., 2022; Ho et al., 2017; Jha & Shah, 2019; Nam et al., 2020). That is, Ho et al. (2017) propose a modelling framework for star rating decisions based on expectations from prior star ratings. Here, taking into account prior textual reviews was out of scope for their research, but they mentioned it as promising future research direction (Ho et al., 2017). Similarly, Nam et al. (2020) analyse disconfirmation resulting from expectations derived from prior reviews, but just based on a binary survey response instead of review texts. Heinrich et al. (2022) analyse long-term dynamics in product ratings based on disconfirmation, but also focusing on prior ratings and not the review texts. Indeed, Jha and Shah (2019) used sentiments in the prior textual reviews for expectations as they consider the textual reviews to have a much deeper social influence than prior star ratings, but only an overall sentiment score instead of an analysis on an aspect-based level. Moreover, they analysed the impact on future review texts instead of star ratings and focused on the mobile phone domain instead of analysing different domains of online consumer reviews. Overall, these existing studies also do not analyse which of the EDT dimensions has the highest unique explanatory power on star ratings. Furthermore, they do not focus on insights for business how to prioritize specific product aspects for product improvements or how to manage consumer expectations on an aspect-level in marketing.

3 Investigating the Explanatory Power of Textual Reviews

In this section, we outline our research model with EDT dimensions derived from textual reviews as well as our hypotheses. Then, we outline our research design to investigate the hypotheses.

3.1 Research Model and Hypotheses

Focusing on online consumer reviews, a central source of information for consumers to establish expectations are prior reviews towards a product (Heinrich et al., 2022; Rohde et al., 2022). Especially prior textual reviews enable consumers to form nuanced expectations (Jha & Shah, 2019; Khan et al., 2022). After product experience, a consumer can then express the perceived performance of various product aspects in a subsequent textual online consumer review. Accordingly, we investigate the EDT dimensions derived from textual reviews in this study. This is considered as a promising approach to obtain detailed insights on consumer evaluations in the literature since, inter alia, it allows to analyse larger datasets compared to consumer surveys and thus can potentially yield more robust results (Nam et al., 2020). Here, existing studies have shown that aspect-based sentiments, which capture consumer sentiments towards different product aspects such as service of restaurants, are vital for analysing textual reviews and can be extracted from large datasets (Binder et al., 2019; Chatterjee et al., 2021). This lays the foundation to investigate the EDT dimensions derived from textual reviews on a sentiment-aspect level. Based on the EDT as theoretical foundation and the identified research gap (cf. Section 2), we thus investigate star ratings of online consumer reviews based on the three EDT dimensions, each derived from textual reviews on a sentiment-aspect-level. This gives our research model as stated in Figure 1.

While existing research suggests that each EDT dimension individually is related to post-purchase evaluations, it is still an open research question whether each EDT dimension actually has a unique explanatory power on post-purchase evaluations (Oliver, 2014; J. Zhang et al., 2022). That is, the literature on the EDT discusses different possible mediation effects between the EDT dimensions such as a “performance and disconfirmation model with fully mediated expectations” (Oliver, 2014).

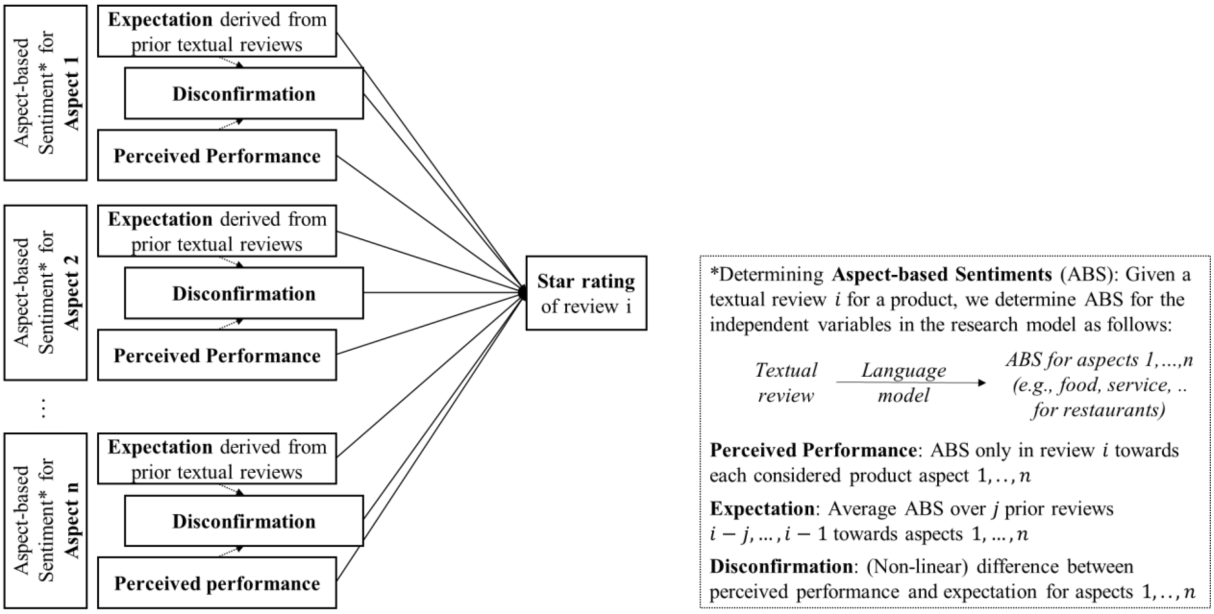

Figure 1. Research model.

To be able to detect whether such full mediation effects exist, an analysis of the unique explanatory power is suited which compares the overall model to models leaving out one EDT dimension. That is, the overall research model with all three EDT dimensions would not have a significantly higher likelihood than a model leaving out a specific EDT dimension when a full mediation effect occurs. This is equivalent to that EDT dimension not having a unique explanatory power for star ratings in our study. Since textual reviews offer detailed information on an aspect-level, we assume that a full mediation effect does not occur between the EDT dimensions derived from textual reviews and that the overall research model (cf. Figure 1) is appropriate here. Thus, our first hypothesis is as follows:

> *H1: Expectation, perceived performance and disconfirmation, each derived from textual online consumer reviews, have a unique explanatory power for star ratings.*

An analysis of the magnitude of the unique explanatory power can also give insights on the extent of mediation effects. For instance, if an EDT dimension has a particularly low unique explanatory power, this indicates that it might be partly mediated by the other EDT dimensions. On the other hand, it is crucial for businesses to have an understanding about which EDT dimension has the highest unique explanatory power to be able to prioritize business measures accordingly for improving the average star rating as key performance indicator (Paget, 2024). Here, *perceived performance* derived from aspect-based sentiments in the review texts has shown to have a strong effect on star ratings in related studies and is the most popular EDT dimension to explain star ratings (cf. Section 2). Thus, our second hypothesis is the following:

> *H2: Perceived performance has a higher unique explanatory power for star ratings than each of both expectations and disconfirmation derived from textual reviews.*

In general, perceived performance and expectations influence disconfirmation (cf. Section 2.1). Especially when modelling disconfirmation based on the perceived performance and expectations derived from textual reviews (cf. Figure 1), the disconfirmation effect might be mediated to a higher extent. Therefore, the unique explanatory power of the EDT dimension *disconfirmation* might be lower compared to *expectations* here. Yet, this might also depend on the type of good that is reviewed. Here, experience goods are goods that generally require direct experience for assessing the performance, while for search goods the required information to assess the product performance is mostly available prior to the purchase (Jin et al., 2023; Mudambi & Schuff, 2010). Accordingly, thorough expectations are typically built especially for search goods, while for experience goods on average less time is spent on searching information beforehand (Mudambi & Schuff, 2010; Ramachandran et al., 2021). Moreover, the disconfirmation effect has shown to be relevant in existing studies on experience goods (Hossain & Quaddus, 2012). Thus, specifically for experience goods, the unique explanatory power of *disconfirmation* derived from textual reviews for star ratings could presumably still be higher than of *expectations*. This gives our third and fourth hypotheses are as follows:

> *H3: Disconfirmation has a higher unique explanatory power for star ratings than expectation derived from textual reviews for experience goods.*

> *H4: Expectation has a higher unique explanatory power for star ratings than disconfirmation derived from textual reviews for search goods.*

3.2 Research Design

To investigate our hypotheses, we use an overall model $M_{overall}^{G}$ for a considered domain G and three partial models $M_{no\_Disconf}^{G}$, $M_{no\_Perf}^{G}$, $M_{no\_Exp}^{G}$ leaving out one EDT dimension respectively. For instance, $M_{no\_Disconf}^{G}$ thus contains independent variables from the dimensions *expectations* and *perceived performance* derived from textual reviews, but not from the EDT dimension *disconfirmation*. Then, we compare the log-likelihood L of all three partial models to the overall model $M_{overall}^{G}$ to investigate whether the left-out EDT dimension has a unique explanatory power on star ratings respectively (Severini, 2000). In this way, we can detect whether a unique explanatory power on star ratings exists rather than considering models based on only one EDT dimension, for which significant results could occur even if an EDT dimension is fully mediated. To that end, we consider different domains G to obtain more robust results.

To assess the unique explanatory power of these EDT dimensions for star ratings, an ordered probit model is used since it has shown to be suited for analysing star ratings based on aspect-based sentiments (Binder et al., 2019). In the following, the integration of the three EDT feature dimensions in this model is outlined in more detail.

According to the ordered probit model, underlying linear preferences $R_*^i \in \mathbb{R}$ can be modelled using the independent variables x_{exp}, x_{perf} and $x_{disconf}$ representing the three EDT dimensions derived from aspect-based sentiments in the textual reviews. This leads to an overall preference model given by

$$R_*^i = \beta_{exp} x_{exp}^i + \beta_{perf} x_{perf}^i + \beta_{disconf} x_{disconf}^i + \epsilon, \tag{1}$$

where $\beta_{exp}\left(= \beta_{\exp 1}, \beta_{exp2}, \dots, \beta_{exp\, n}\right)$, β_{perf} and $\beta_{disconf}$ denote the parameters with respect to the independent variables $x_{exp}^i\left(= x_{exp1}^i, x_{exp2}^i, \dots, x_{exp\, n}^i\right)$, x_{perf}^i and $x_{disconf}^i$ for the n different product aspects in the i-th review, and $\epsilon \sim N(0,1)$ denotes the random error term.

Then, a discrete random variable $R^i \in \{1, \dots, 5\}$ and thresholds $\theta_1, \dots, \theta_4$ are used to estimate the actual star rating $r^i \in \{1, \dots, 5\}$ in the review i, which means,

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

This model with the parameters $\beta_{exp}, \beta_{perf}$ and $\beta_{disconf}$ as well as the thresholds $\theta_1, \dots, \theta_4$ (cf. Equations (1) and (2)) constitutes our overall research model $M_{overall}^{G}$ for a domain G.

To determine the independent variables x_{exp}^i, x_{perf}^i and $x_{disconf}^i$, we use aspect-based sentiments (ABS). For instance, in the exemplified review sentence "The waiter was attentive and friendly." first the aspect term "waiter" is extracted by a fine-tuned language model and assigned to the aspect service (Xu et al., 2019). Subsequently, another fine-tuned language model assigns a positive sentiment towards service based on the term "attentive and friendly". These tasks are referred to as aspect term extraction and aspect-sentiment classification in research on aspect-based sentiment analysis (Xu et al., 2019). If there are multiple sentiments towards a specific aspect expressed in a review i (cf. Appendix C for an exemplified review), we averaged the sentiments towards this aspect to determine x_{perf}^i. That is, for the EDT dimension *perceived performance* we consider the aspect-based sentiments in the review i for n product aspects. Here, all sentiments not addressing a considered product aspect are subsumed under an aspect 'miscellaneous' to comprehensively account for the expressed sentiments. For the EDT dimension *expectations*, we use the average aspect-based sentiments based on the latest $j = 20$ reviews for the considered product prior to the review i, since consumers typically do not read the text of more than 20 reviews and the latest reviews are especially relevant (Askalidis et al., 2017; F. Wang et al., 2018). As outlined above, this is promising since prior textual reviews are a key factor in shaping consumers' expectations (Jha & Shah, 2019). For the EDT dimension *disconfirmation*, we consider the cubic difference between expectations and perceived performance on a sentiment-aspect-level since this allows to take into account large deviations which are considered especially relevant for disconfirmation (Brown et al., 2014). That is, existing research found that using a cubic model was the best to explain the nuances of the complex relationship between expectations and perceived performance and thus adequately represent disconfirmation (Brown et al., 2014). To investigate our hypotheses for textual online consumer reviews, we accordingly focus on disconfirmation based on expectations derived from prior textual reviews in this study.

4 Data and Results

We start this section by introducing the selected datasets and their preparation. Based on that, we present and discuss the results with respect to our four hypotheses.

4.1 Datasets and Data Preparation

To analyse our hypotheses, we used three datasets from the experience good domains of restaurants and movies as well as the search good domain of laptops. Reviews of these three domains are also commonly used for evaluating approaches for sentiment analysis (Thet et al., 2010; Zheng & Li, 2024). More precisely, movies also contain some search good attributes such as movie trailers and thus are not pure experience goods. That is, these three domains show variability along the established distinction between experience goods and search goods (Ramachandran et al., 2021). Therefore, they are suited to analyse H3 and H4 in detail as well as to enhance the robustness of our results in general.

The restaurant dataset consists of reviews for restaurants, bars and cafés in New York City from an established web portal for local businesses. The movie dataset consists of reviews for movies and other video content (e.g., documentaries, recorded concerts, etc.), while the laptop dataset consists of reviews for laptops, notebooks and tablet computers, both originating from the Amazon review dataset provided by Ni et al. (2019). Thereby, each review in the above datasets consists of a textual consumer review with an associated star rating on a five-tier scale from 1 star to 5 stars. Further, each dataset exhibits the widely recognized "J-shaped" rating distribution (e.g., Debortoli et al., 2016). Each dataset is large enough to analyse various independent variables of different EDT dimensions in an explanatory model. That is, the number of events per independent variable (EPV) is higher than 1,000 in our analyses as the smallest dataset has 28,000 reviews and we consider less than 28 independent variables. Here, a value of 100 is already considered as sufficient in the literature (e.g., Heinze et al., 2018). Moreover, each dataset contains more than 100 different products, whereby the laptop dataset has the highest amount due to individual listings of specific product variations. An overview of the datasets is given in Table 1.

Dataset	**Movies**	**Laptops**	**Restaurants**
# of reviews	35k	28k	32k
Rating distribution, i.e., (relative) # of ratings per level of star rating	1 star: 9% (~3.1k) 2 stars: 6% (~ 2.0k) 3 stars: 9% (~3.0k) 4 stars: 18% (~6.2k) 5 stars: 59% (~21k)	1 star: 13% (~3.7k) 2 stars: 6% (~ 1.6k) 3 stars: 8% (~2.3k) 4 stars: 21% (~5.8k) 5 stars: 52% (~15k)	1 star: 6% (~1.8k) 2 stars: 6% (~2.0k) 3 stars: 13% (~4.1k) 4 stars: 35% (~11.1k) 5 stars: 40% (~12.8k)
# of consumers # of products	33k 166	24k 757	27k 185
Avg. review length (characters)	463	810	619

Table 1. Description of the datasets.

To operationalise our research model and capture consumer sentiments for different product aspects, we extracted aspect-based sentiments from the review texts as outlined in the following. We used a supervised aspect extraction method based on the large language model BERT (Devlin et al., 2019), since large language models are the state of the art method for aspect extraction from text (J. Yang et al., 2023). Here, an encoder model such as BERT – compared to decoder-only models such as GPT – is especially suited for aspect extraction tasks (Zheng & Li, 2024). This allows for directly comprehensible aspects of high validity (Devlin et al., 2019) and thus strengthens the reliability of the results. Notably, this also constitutes a novelty compared to the related work using unsupervised topic modelling approaches that have to be interpreted manually (Jabr et al., 2018) or lexicon-based extraction techniques (Chatterjee et al., 2021) that generally achieve much lower validity compared to large language models such as BERT (Devlin et al., 2019). To further improve the quality of the aspect extraction, we used the post-trained BERT models for each specific domain of our considered datasets (Xu et al., 2019), since the post-trained BERT models have a stronger alignment to the domain-specific use of language. In more detail, we fine-tuned these models for aspect term extraction and aspect-sentiment classification (Sun et al., 2019). To assess the quality of the data preparation, we first analysed the aspect term extraction conducted by the fine-tuned BERT model. Here, F1 scores of 0.78, 0.75 and 0.76 for restaurants, movies and laptops were achieved. Based on this extraction, the aspect-based sentiment classification yielded F1 scores of 0.84, 0.86 and 0.80, respectively. All F1 scores are comparable to the state-of-the-art (Xu et al., 2019; Zheng & Li, 2024). That is, using the encoder model BERT as a basis achieves top-performing results for these tasks (Zheng & Li, 2024).

An overview of the considered variables for each EDT dimension and dataset is given in Table 3. For each dataset, we considered five aspects in accordance with previous studies on aspect-based sentiment analysis for these domains (Binder et al., 2019; Thet et al., 2010; J. Wang et al., 2018). Sentiments that do not refer to these five aspects were subsumed as a sixth miscellaneous sentiment. Finally, to verify the stability of our research model, multicollinearity between the independent variables was measured by the variance inflation factor (VIF). The maximum VIFs were 3.18, 2.99 and 2.97 for movies, laptops and restaurants respectively, whereby VIF values less than 10 are uncritical regarding model stability (O'brien, 2007). Based on the extracted aspect-based sentiments for each review *i* in the datasets, we then determined the independent variables x_{exp}^{i}, x_{perf}^{i} and $x_{disconf}^{i}$ as outlined in Section 3.2.

4.2 Results

To investigate our hypotheses, we conducted different log-likelihood tests as outlined in Table 2. Here, the three log-likelihood tests for H1 all give a significantly higher likelihood of the overall research model $M_{overall}^{G}$ for all three considered domains, i.e. H1 is supported. For H2, the two log-likelihood tests each give a significantly lower likelihood of the model $M_{no\_Perf}^{G}$, i.e., H2 is also supported. That is, our results consistently support that each of the three EDT dimensions *expectations, perceived performance* and *disconfirmation* derived from textual reviews have a unique explanatory power on star ratings with perceived performance having the highest unique explanatory power.

Further, the log-likelihood test for H3 does not give a significantly higher likelihood of $M_{no\_Exp}^{G}$ compared to $M_{no\_Disconf}^{G}$, i.e., *disconfirmation* does not have a significantly higher unique explanatory power than *expectations* for the experience good $G = restaurants$ nor for $G = movies$. Thus, H3 is not supported. For H4, $M_{no\_Disconf}^{G}$ does have a significantly higher likelihood than $M_{no\_Exp}^{G}$ for the search good $G = laptops$, i.e., H4 is supported. Therefore, our results do not support the presumed differentiation between search goods and experience goods but suggest that the EDT dimension *expectations* derived from prior textual reviews consistently has a higher unique explanatory power for star ratings than *disconfirmation*. The results of our hypotheses tests are summarized in Table 2. The coefficients of the research model for the three considered domains restaurants, laptops and movies are given in Table 3. Based on that, our results can be discussed on a detailed aspect-level.

Hypothesis	Overview of log-likelihood ratio tests for each hypothesis	p-value *(results were consistent for all considered domains)*	Result	Considered domains *G*
H1	$L(M^G_{overall}) > L(M^G_{no\_Perf})$ $L(M^G_{overall}) > L(M^G_{no\_Exp})$ $L(M^G_{overall}) > L(M^G_{no\_Disconf})$	*** (p<0.001) *** (p<0.001) *** (p<0.001)	Supported	Movies, laptops, restaurants
H2	$L(M^G_{no\_Perf}) < L(M^G_{no\_Exp})$ $L(M^G_{no\_Perf}) < L(M^G_{no\_Disconf})$	*** (p<0.001) *** (p<0.001)	Supported	Movies, laptops, restaurants
H3	$L(M^G_{no\_Exp}) > L(M^G_{no\_Disconf})$ for experience goods	-	Not supported	Restaurants, movies
H4	$L(M^G_{no\_Disconf}) > L(M^G_{no\_Exp})$ for search goods	*** (p<0.001)	Supported	Laptops

Table 2. Results of hypotheses tests.

	Ind. Variable Movie	Movie Coefficient	Ind. Variable Laptop	Laptop Coefficient	Ind. Variable Restaurant	Restaurant Coefficient
Perceived Performance	story_sent	0,15 ***	performance_sent	0,48 ***	food_sent	0,61 ***
	cinematography_sent	0,21 ***	support_sent	0,18 ***	service_sent	0,23 ***
	acting_sent	0,16 ***	screen_sent	0,27 ***	ambience_sent	0,07 ***
	music_sent	0,05 ***	battery_sent	0,11 ***	quantity_sent	0,06 ***
	price_sent	0,07 ***	price_sent	0,04 ***	price_sent	0,08 ***
	misc_sent	0,42 ***	misc_sent	0,51 ***	misc_sent	0,5 ***
Expectation	avg_story_sent_prev	0,1 ***	avg_performance_sent_prev	0,09 ***	avg_food_sent_prev	0,1 ***
	avg_cinematography_sent_prev	0,11 ***	avg_support_sent_prev	-0,04 ***	avg_service_sent_prev	0,11 ***
	avg_acting_sent_prev	0,06 ***	avg_screen_sent_prev	0,04 ***	avg_ambience_sent_prev	-0,03 ***
	avg_music_sent_prev	-0,02	avg_battery_sent_prev	0,11 ***	avg_quantity_sent_prev	0
	avg_price_sent_prev	0,1 ***	avg_price_sent_prev	-0,08 ***	avg_price_sent_prev	0,03 **
	avg_misc_sent_prev	0,12 ***	avg_misc_sent_prev	0,07 ***	avg_misc_sent_prev	0,05 ***
Disconfirmation	disconf_story	0,06 ***	disconf_performance	0,05 ***	disconf_food	-0,05 ***
	disconf_cinematography	0,04 ***	disconf_support	-0,02	disconf_service	0,1 ***
	disconf_acting	-0,01	disconf_screen	0,05 ***	disconf_ambience	-0,01
	disconf_music	-0,03 **	disconf_battery	0,06 ***	disconf_quantity	-0,01
	disconf_price	0,05 ***	disconf_price	-0,03 **	disconf_price	-0,01
	disconf_misc	0,1 ***	disconf_misc	0,1 ***	disconf_misc	-0,01

*Table 3. Coefficients in the research model (***: p<0.001; **: p<0.01; *: p<0.05).*

5 Discussion and Implications

In this section, we discuss our results with respect to the four hypotheses. Based on that, we outline implications of our study for research and practice.

5.1 Discussion

Our results consistently support that all three EDT dimensions have a unique explanatory power on star ratings with perceived performance having the highest one. In more detail, our research model on a sentiment-aspect-level allows to analyse the coefficients for the different aspects. For instance, besides the hypothesis tests supporting H2, we also see that for the EDT dimension *perceived performance* all considered aspects have significant positive coefficients for each domain (cf. Table 3) and can be directly interpreted. This supports that for all considered aspects in each domain its perceived performance generally has a significant effect on star ratings and strengthens the robustness of our result for H2. Here, our research model also accounts for expectations and disconfirmation derived from textual reviews and thus gives more comprehensive insights than studies focusing solely on the perceived performance (Ho et al., 2017). In general, our hypotheses are evaluated on larger datasets of online reviews compared to, e.g., consumer surveys, which further strengthens the robustness of our results (Nam et al., 2020).

Differences between the domains can be observed in a detailed analysis. To that end, we use the Nagelkerke pseudo R-squared to quantitatively assess the explanatory power of the research model (cf. Appendix A). Based on an analysis of partial R-squared values, we can assess that *perceived performance* has a clearly lower unique explanatory power for movies compared to restaurants and laptops in our research model (24.7% vs 47.6% and 46.0% respectively, cf. Appendix B). Here, it can also be seen that expectations derived from prior textual reviews have a larger unique explanatory power for movies compared to restaurants and laptops (cf. Appendix B). We discuss this in more detail in the following:

Expectations derived from prior textual reviews consistently constitute the EDT dimension with the second highest unique explanatory power in the research model (cf. Section 4.2 and Appendix B). Yet, our results do not support that the unique explanatory power of expectations derived from prior textual reviews clearly differs between search goods and experience goods, but rather for specific goods such as movies. That is, *perceived performance* has a relatively low unique explanatory power for movies, while expectations derived from prior textual reviews have a comparably high unique explanatory power here (cf. Appendix B). This might be related to existing findings which observed that consumer personality traits have a particularly high explanatory power for star ratings of movies (Karumur et al., 2016). That is, star ratings for movies seem to be more influenced by additional factors such as personality traits of consumers. This can be analysed in more detail by means of our research model and recent approaches for extracting Big5 personality traits from review texts (K. Yang et al., 2023).

Overall, expectations derived from prior textual reviews indeed have a positive effect on star ratings for most aspects as suggested by the EDT (cf. mostly positive coefficients in Table 3). An exception is given, e.g., by the price and service expectation of laptops (cf. coefficients of $avg\_price\_sent\_prev = -0.08$ and $avg\_service\_sent\_prev = -0.04$ for laptops). This could be connected to initial offerings such as discounts or free returns for new laptop models that get cancelled later and potentially lead to disappointed price or service expectations.

Moreover, this study gives more detailed insights on the established positive impact of positive disconfirmation on star ratings in the literature (Nam et al., 2020). First, the largest coefficients for the EDT dimension disconfirmation are significantly positive in our research model for each domain, which is consistent to existing research (cf. $disconf\_service = 0.10$*** for restaurants, $disconf\_battery = 0.06$*** for laptops or $disconf\_story = 0.06$*** for movies). Here, our results further suggest that disconfirmation is especially relevant for main aspects of a product. That is, the results indicate that only disconfirmation towards main aspects of a product leads to delight or disappointment with corresponding disproportionate weighting in the star rating assignment. For instance, while for restaurants an unexpectedly unfriendly service alone can lead to a low star rating (cf. $disconf\_service = 0.10$***), this is typically not the case, e.g., solely due to an underwhelming ambience (cf. $disconf\_ambience = -0.01$).

5.2 Implications

Existing literature already started to investigate how consumers' star ratings decisions are influenced by expectations derived from prior star ratings (Guo & Zhou, 2016; Nam et al., 2020). The existing body of knowledge in this research area is extended by deriving the EDT dimensions *expectations, perceived performance* and *disconfirmation* from textual reviews on a sentiment-aspect-level in this study. For instance, our results suggest that large differentiations between expectations and perceived performance are especially relevant for specific aspects of a product (cf. Table 3). Such detailed insights cannot be obtained by considering only star ratings rather than the textual reviews.

Moreover, our results for H1 support that all three EDT dimensions derived from textual reviews have a unique explanatory power on star ratings, i.e., are not fully mediated. This constitutes a novel insight compared to, e.g., existing modelling frameworks deriving expectations from prior ratings that presume expectations to be fully mediated (Ho et al., 2017). This suggests that when deriving the EDT dimensions from textual reviews, all three EDT dimensions should be considered to have a direct effect on star ratings. Yet, the EDT dimension disconfirmation has a rather low unique explanatory power in our research model. Thus, future research using our study as a basis for modelling frameworks based on textual reviews can investigate possible partial mediation effects for disconfirmation in detail and focus on disconfirmation towards main aspects of a product.

Overall, this study can also help to explain findings in related studies which observed non-linear effects of sentiments towards specific product aspects on post-purchase evaluations (Shin & Nicolau, 2022). In our analyses, we can investigate such non-linear effects through capturing disconfirmation by means of a cubic term to explicitly account for strong deviations from expectations. By applying this cubic term, also the „zone of indifference“ for disconfirmation – which is discussed in existing research on the EDT – is taken into account, but on a more detailed sentiment-aspect level instead of prior star ratings as in existing research (Brown et al., 2014; Oliver, 2014).

In this study, we focus on expectations and disconfirmation based on prior textual reviews. Here, our research model can also be used to investigate whether these findings translate to different sources of expectations (Nam et al., 2020). That is, whether using, e.g., assessments of movie trailers or laptop ads, to determine expectations rather than prior textual reviews also yields similar results for different product aspects and gives a higher unique explanatory power of expectations for movies in our research model. Furthermore, our study also has theoretical implications on the research area of repurchase intentions. That is, large positive disconfirmation, sometimes referred to as delight, is a crucial factor for repurchase intention (Oliver, 2014). This can be analysed in more detail based on aspect-based sentiments by means of our research model. Since our results suggest that disconfirmation towards specific aspects is especially relevant for star ratings, these aspects might also be especially relevant for repurchase intention.

In general, analyses of textual online consumer reviews are considered as especially promising in the eWOM research field (Donthu et al., 2021). We are confident that this study encourages further research on textual online consumer reviews. This is particularly feasible for eWOM research due to the recent availability of large language models. Moreover, our research model contributes to a careful understanding of influence factors for consumers' expectation, perceived performance and disconfirmation. That is, while existing research on online consumer reviews predominantly used prior star ratings as a basis, by means of our research model expectations, perceived performance and corresponding disconfirmation towards different product aspects are taken into account. In general, a research gap for textual analyses has been detected since most approaches focus only on the extraction of one kind of textual meaning such as an overall sentiment score of a review (Baden et al., 2022). This is considered an issue since the information required to address specific research questions is typically scattered unevenly across the text (Baden et al., 2022). We contribute to that gap in our study by investigating the EDT dimensions derived from textual reviews on a sentiment-aspect level.

Our results on a sentiment-aspect-level give actionable insights for practitioners without the need of time-consuming and costly consumer surveys (Chatterjee et al., 2021). In more detail, our insights on the impact of different product aspects on star ratings can help to prioritize business measures for product improvements (Chatterjee et al., 2021). In that way, this study provides insights for managers to efficiently increase the average star rating as key performance indicator for businesses (Jabr & Rahman, 2022).

Moreover, our insights on the unique explanatory power of expectations and disconfirmation for star ratings can support businesses to manage consumer expectations. For instance, this study can support the development of specific marketing strategies such as presenting a mix of positive and negative reviews to consumers (Jha & Shah, 2021). That is, marketers can analyse whether it is beneficial to lower expectations for specific product aspects to reduce the risk of negative disconfirmation (L. Zhang et al., 2021). Here, the coefficients in the research model can be taken into account to determine product aspects where expectations and disconfirmation are particularly relevant for consumer satisfaction (e.g., service quality of a restaurant). Our results suggest that for domains such as movies it can be an especially relevant marketing strategy to steer positive textual reviews after launch, e.g., by actively inviting the first consumers to leave a textual review. Overall, this study thus supports marketers to better understand and address the needs of the consumers (Dhar & Bose, 2022).

6 Conclusion

Both research and practice have acknowledged the importance of analysing star ratings. Based upon the existing body of knowledge on this topic, our work is the first to investigate the EDT based on expectations derived from prior textual reviews on a sentiment-aspect-level in online consumer reviews. To investigate four hypotheses, we operationalized our research model on three review datasets using the large language model BERT. Our results support, inter alia, that all three EDT dimensions expectations, perceived performance and disconfirmation, each derived from textual reviews, have a unique explanatory power for star ratings. Moreover, we find support the EDT dimension perceived performance derived from textual reviews has the highest unique explanatory power on star ratings here.

Nevertheless, the work at hand has some limitations which could be a starting point for further research. First, we discussed the domain-specific results for our third and fourth hypothesis by means of considering search goods and experience goods. To that end, analysing our research model on further domains could substantiate and broaden these investigations. Also, detailed analyses of specific consumer groups could constitute interesting complementary analyses to our work. Finally, based on this study it could be analysed how consumer personality traits and consumer contexts moderate the impact of disconfirmation derived from textual reviews on star ratings in online consumer reviews.

References

Anderson-Sprecher, R. (1994). Model comparisons and R 2. *The American Statistician*, *48*(2), 113–117.

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

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16*(1), 1–18.

Binder, M., Heinrich, B., Klier, M., Obermeier, A., & Schiller, A. (2019). Explaining The Stars: Aspect-based Sentiment Analysis of Online Customer Reviews. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Brown, S. A., Venkatesh, V., & Goyal, S. (2014). Expectation confirmation in information systems research. *MIS Quarterly, 38*(3), 729-A9.

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.

Debortoli, S., Müller, O., Junglas, I., & Vom Brocke, J. (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, *39*, 110–135. https://doi.org/10.17705/1CAIS.03907

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*.

Dhar, S., & Bose, I. (2022). Walking on air or hopping mad? Understanding the impact of emotions, sentiments and reactions on ratings in online customer reviews of mobile apps. *Decision Support Systems*, *162*, 113769.

Donthu, N., Kumar, S., Pandey, N [Neeraj], Pandey, N [Nitesh], & Mishra, A. (2021). Mapping the electronic word-of-mouth (eWOM) research: A systematic review and bibliometric analysis. *Journal of Business Research*, *135*, 758–773.

Guo, B., & Zhou, S. (2016). Understanding the impact of prior reviews on subsequent reviews: The role of rating volume, variance and reviewer characteristics. *Electronic Commerce Research and Applications*, *20*, 147–158.

Gutt, D., Neumann, J., Zimmermann, S., Kundisch, D., & Chen, J. (2019). Design of review systems - A strategic instrument to shape online reviewing behavior and economic outcomes. *The Journal of Strategic Information Systems*, *28*(2), 104–117.

Han, S., & Anderson, C. K. (2020). Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, *61*(2), 142–153.

Heinrich, B., Hollnberger, T., Hopf, M., & Schiller, A. (2022). Long-term Sequential and Temporal Dynamics in Online Consumer Ratings. In *ECIS 2022 Research Papers.*

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection-a review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449.

Ho, Y.-C., Wu, J., & Tan, Y. (2017). Disconfirmation effect on online rating behavior: A structural model. *Information Systems Research*, *28*(3), 626–642.

Hossain, M. A., & Quaddus, M. (2012). Expectation-confirmation theory in information system research: A review and analysis. *Information Systems Theory*, 441–469.

Hult, G. T. M., Morgeson III, F. V., Sharma, U., & Fornell, C. (2022). Customer satisfaction and international business: A multidisciplinary review and avenues for research. *Journal of International Business Studies*, *53*(8), 1695–1733.

Jabr, W., Cheng, Y [Yichen], Zhao, K., & Srivastava, S. (2018). What Are They Saying? A Methodology for Extracting Information from Online Reviews. In *Proceedings of the 39th International Conference on Information Systems*.

Jabr, W., Liu, B [Ben], Yin, D., & Zhang, H. (2020). Research Curation on Online Word-of-Mouth. *MIS Quarterly*.

Jabr, W., & Rahman, M. S. (2022). Online Reviews and the Information Overload: The Role of Selective, Parsimonious and Concordant Top Reviews. *MIS Quarterly*, *46*(3), 1517–1550.

Jha, A. K., & Shah, S. (2019). Social Influence on Future Review Sentiments: An Appraisal-Theoretic View. *Journal of Management Information Systems*, *36*(2), 610–638.

Jha, A. K., & Shah, S. (2021). Disconfirmation effect on online review credibility: An experimental analysis. *Decision Support Systems*, *145*, 113519.

Jin, J., Lin, C., Wang, F [Fenghua], Xu, T., & Zhang, W. (2023). A study of cognitive effort involved in the framing effect of summary descriptions of online product reviews for search vs. experience products. *Electronic Commerce Research*, *23*(2), 785–806.

Karumur, R. P., Nguyen, T. T., & Konstan, J. A. (2016). Exploring the value of personality in predicting rating behaviors: A study of category preferences on MovieLens. In *Proceedings of the 10th ACM conference on recommender systems*.

Khan, F. M., Khan, S. A., Shamim, K., Gupta, Y., & Sherwani, S. I. (2022). Analysing customers' reviews and ratings for online food deliveries: A text mining approach. *International Journal of Consumer Studies*.

Kolomoyets, Y., & Dickinger, A. (2023). Understanding value perceptions and propositions: A machine learning approach. *Journal of Business Research*, *154*, 113355.

Legendre, P., & Legendre, L. F. J. (2012). *Numerical ecology* (3rd Edition, Vol. 24). Elsevier.

Liu, H., Jayawardhena, C., Shukla, P., Osburg, V.-S., & Yoganathan, V. (2024). Electronic word of mouth 2.0 (eWOM 2.0)-The evolution of eWOM research in the new age. *Journal of Business Research*, *176*, 114587.

Ma, X., Khansa, L., Deng, Y., & Kim, S. S. (2013). Impact of prior reviews on the subsequent review process in reputation systems. *Journal of Management Information Systems*, *30*(3), 279–310.

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

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

Nam, K., Baker, J., Ahmad, N., & Goo, J. (2020). Determinants of writing positive and negative electronic word-of-mouth: Empirical evidence for two types of expectation confirmation. *Decision Support Systems*, *129*, 113168.

Ni, J., Li, J [Jiacheng], & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 188–197). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1018

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Oliver, R. L. (2014). *Satisfaction: A behavioral perspective on the consumer* (2nd edition). Routledge.

Paget, S. (2024). *Local Consumer Review Survey 2024*. https://www.brightlocal.com/research/local-consumer-review-survey/

Qahri-Saremi, H., & Montazemi, A. R. (2022). Negativity bias in the diagnosticity of online review content: the effects of consumers' prior experience and need for cognition. *European Journal of Information Systems*, 1–18.

Ramachandran, R., Sudhir, S., & Unnithan, A. B. (2021). Exploring the relationship between emotionality and product star ratings in online reviews. *IIMB Management Review*, *33*(4), 299–308.

Ren, J., & Nickerson, J. V. (2019). Arousal, valence, and volume: how the influence of online review characteristics differs with respect to utilitarian and hedonic products. *European Journal of Information Systems*, *28*(3), 272–290.

Rohde, C., Kupfer, A., & Zimmermann, S. (2022). Explaining reviewing effort: Existing reviews as potential driver. *Electronic Markets*, *32*(3), 1169–1185.

Rubin, A. (2024). *Yelp 20th Anniversary*. https://blog.yelp.com/news/yelp-20th-anniversary/

Sahoo, N., Dellarocas, C., & Srinivasan, S. (2018). The impact of online product reviews on product returns. *Information Systems Research*, *29*(3), 723–738.

Severini, T. A. (2000). *Likelihood methods in statistics*. Oxford University Press.

Shin, S., & Nicolau, J. L. (2022). Identifying attributes of wineries that increase visitor satisfaction and dissatisfaction: Applying an aspect extraction approach to online reviews. *Tourism Management*, *91*, 104528.

Siering, M., & Janze, C. (2019). Information Processing on Online Review Platforms. *Journal of Management Information Systems*, *36*(4), 1347–1377.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics.* Symposium conducted at the meeting of Springer.

Thet, T. T., Na, J.-C., & Khoo, C. S. G. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, *36*(6), 823–848.

Wang, F [Fang], Menon, K., & Ranaweera, C. (2018). Dynamic trends in online product ratings: A diagnostic utility explanation. *Journal of Business Research*, *87*, 80–89.

Wang, J., Li, J [Jie], Li, S., Kang, Y., Zhang, M., Si, L., & Zhou, G. (2018). Aspect sentiment classification with both word-level and clause-level attention networks. In *International Joint Conference on Artificial Intelligence*.

Xu, H., Liu, B [Bing], Shu, L., & Yu, P. S. (2019). BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ArXiv Preprint ArXiv:2304.13712*.

Yang, K., Lau, R. Y. K., & Abbasi, A. (2023). Getting Personal: A Deep Learning Artifact for Text-Based Measurement of Personality. *Information Systems Research*, *34*(1), 194–222.

Zhang, J., Chen, W., Petrovsky, N., & Walker, R. M. (2022). The expectancy-disconfirmation model and citizen satisfaction with public services: A meta-analysis and an agenda for best practice. *Public Administration Review*, *82*(1), 147–159.

Zhang, L., Wei, W., Line, N. D., & Cheng, Y [Yusi] (2021). When positive reviews backfire: The effect of review dispersion and expectation disconfirmation on Airbnb guests' experiences. *International Journal of Hospitality Management*, *96*, 102979.

Zheng, Y., & Li, X. (2024). You Only Read Once: Constituency-Oriented Relational Graph Convolutional Network for Multi-Aspect Multi-Sentiment Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Appendix A

To assess the explanatory power of the research model $M_{overall}^{G}$ for star ratings, the Nagelkerke pseudo R-squared is used. By this measure, the likelihood of the research model is compared to a null-model (Maddala, 1983), which does not take the independent variables into account. That is, the null-model does not distinguish between different reviews, but still determines the thresholds $\theta_1, .., \theta_4$ according to the rating distribution in the dataset. In detail, the used comparison of likelihoods is equal to the common R-squared measure in case of a linear regression. Overall, the Nagelkerke pseudo R-squared is given by

$$\mathcal{R}_{Nagelkerke}^{2} = \frac{1 - \left[\frac{L_{Null-Model}^{G}}{L_{M_{overall}}^{G}}\right]^{2/N}}{1 - L_{Null-Model}^{G}{}^{2/N}}, \qquad (4)$$

where $L_{M_{overall}^{G}}$ and $L_{Null-Model}^{G}$ denote the value of the likelihood function at the maximum likelihood estimate of the research model and the null-model respectively (Binder et al., 2019). Further, N denotes the number of reviews in the dataset here. Thereby, the range $[0,1]$ (i.e., $[0\%, 100\%]$) of the Nagelkerke pseudo R-squared measure is in accordance with the common R-squared measure for linear regression models.

Appendix B

To investigate the explanatory power of the research model in more detail, we evaluated partial R-squared values (Anderson-Sprecher, 1994). That is, we determined how much explanatory power is lost by leaving out a specific EDT dimension (i.e., by comparing $M_{overall}^{G}$ to the partial models as introduced in Section 3.2). To directly compare the results to the explanatory power (e.g., Nagelkerke R-squared of 54.0% for the restaurant domain), we assessed the unique explanatory power of the each EDT dimension by scaling the partial Nagelkerke R-squared values to this benchmark (e.g., cf. Legendre & Legendre, 2012). The results of this additional analysis are given in Table 4.

Dataset	**EDT dimension**		
	Perceived Performance	**Expectation**	**Disconfirmation**
Movies (39.2% in sum)	24.7%	12.8%	1.6%
Laptops (53.7% in sum)	46.0%	6.3%	1.4%
Restaurants (54.0% in sum)	47.6%	5.5%	0.9%

Table 4. Unique explanatory power for star ratings of each EDT dimension in the research model measured by scaled partial Nagelkerke R-squared.

Appendix C

In the following, an exemplary review with average length in the restaurant dataset is given:

Finally made it to Jane for brunch. We were late for our reservations, and the front staff was very accommodating. It was very late in the day, and the place was still humming. We sat comfortably in the back along the banquet. Anytime is the right time for eggs, so we went for the items on the left side of the menu that came with a complimentary drink. I opted for the mimosa (meh, but i don't usually go that route) and my friend got a Bloody Mary which she happily drank, but looked a bit paltry in my eyes. My companion opted for the eggs with black beans. The waiter was attentive and friendly.