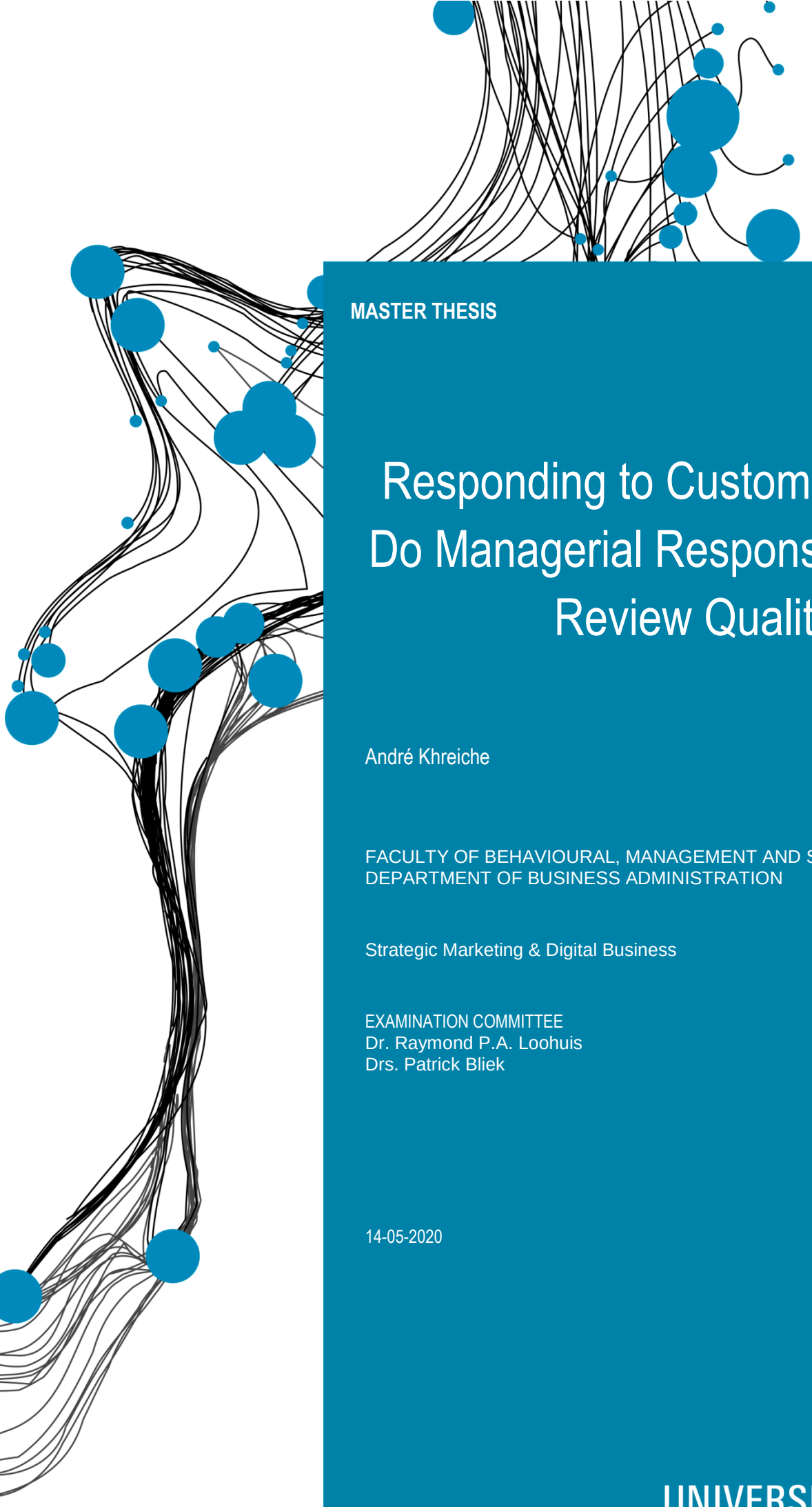

MASTER THESIS

Responding to Customer Reviews: Do Managerial Responses Enhance Review Quality?

André Khreiche

FACULTY OF BEHAVIOURAL, MANAGEMENT AND SOCIAL SCIENCES
DEPARTMENT OF BUSINESS ADMINISTRATION

Strategic Marketing & Digital Business

EXAMINATION COMMITTEE
Dr. Raymond P.A. Loohuis
Drs. Patrick Bliek

14-05-2020

UNIVERSITY OF TWENTE.

Acknowledgements

This thesis marks the completion of my Master's programme in Business Administration at the University of Twente. Several people have helped me during the time it took to write it and during the years of education leading up to it. I would like to take this opportunity to express my gratitude to them.

I would like to thank my family and friends for their support throughout my education. First and foremost, I am deeply grateful to my parents, Abdallah Khreiche and Ingrid Khreiche as well as my siblings, Patricia Khreiche and Mario Khreiche. Furthermore, I am thankful to all my good friends from many different places, whose companionship I've been enjoying greatly for the past years.

I am also extremely grateful to my supervisors, who were always helpful, reliable and inspiring. Dr. Raymond P.A. Loohuis was dedicated in helping me improve this thesis and always provided excellent and encouraging suggestions. Drs. Patrick Bliek, whose time and effort is highly valued, agreed to step in as my second supervisor. Finally, Dr. Anna Priante, who supervised my thesis throughout the early stages, gave a lot of indispensible input and helped to set the course for a successful project.

Management Summary

In the course of the last two decades, online customer reviews have become an instrumental source of information for customers in guiding their purchasing process. Similarly, firms have embraced online reviews as a channel for marketing, customer care and public relations management. Many reviewing platforms, such as TripAdvisor or Trustpilot, offer organizations the option to reply to reviews of their customers. However, the degree to which this feature is used by companies varies greatly. I investigate whether higher response frequency, higher response speed and higher response length, are effective in increasing the quality of online reviews on the review platform Trustpilot. By combining traditional review quality metrics with metrics based on natural language processing and machine learning techniques, this study employs a nuanced view of online review quality. The analysis reveals that higher response frequency is not effective in improving the quality of online reviews. Both, higher response speed and higher response length, on the other hand, are shown to improve the quality of online reviews. These findings bear implications for companies which engage in responding to customer reviews. Company representatives should be selective in deciding which reviews to respond to, rather than aiming to respond to as many reviews as possible during the time dedicated to this task. It is advisable to respond in a manner which best informs the potential audience of the conversation. This includes the individual who posted the review as well as other (potential) customers. Furthermore, the task of responding to customer reviews should be performed regularly and in short intervals, rather than waiting for reviews to accumulate and answering them in bulk, so that the time span between review and response can be minimized.

Keywords:
Online Reviews, Customer Reviews, Review Platforms, Managerial Response, Webcare, Review Quality, Helpfulness, Readability, Diagnosticity, Machine Learning,

Contents

1 Introduction

1.1 Situation & Complication

Online customer reviews have become increasingly important for consumers and organizations in a broad range of industries. Numerous online platforms such as Amazon, TripAdvisor, Yelp or Trustpilot provide users easy public access to customer evaluations of companies, products and services. For companies, these online platforms can serve as a means of communication, marketing and customer care (Luca & Zervas, 2016) which can be instrumental to businesses' reputation and success (Zhang & Vásquez, 2014). On some online review platforms, including Trustpilot, TripAdvisor or Expedia companies have the option to publicly respond to the reviews posted by consumers to, for instance, express gratitude for the feedback or address issues and complaints. However, the degree to which this option is embraced by companies varies greatly. While some firms do not answer any reviews, others reply to a large share of reviews and do so in great detail. A firm's strategy in responding to online reviews may impact the quality of subsequent reviews. There is a lot of research suggesting that managerial responses (MRs), or the lack thereof, affect consumer perceptions and behavior. For instance, in a TripAdvisor survey[1], 77% of the respondents indicated that they were more likely to book a hotel which responds to customer reviews. MRs do not just directly influence the customer who receives the response, readers of the conversation between the reviewer and the responding organization are also affected in their decision making process (Chen, Gu, Ye, & Zhu, 2019). This externality places additional importance on managerial responses.

1.2 Research Gap & Purpose

In many studies, positive effects of MRs on financial performance (Lui, Bartosiak, Piccoli, & Sadhya, 2018), review quantity (Proserpio & Zervas, 2017) and rating score (C. Li, Cui, & Peng, 2017) have been reported. However, an aspect for which research is far more scant is the effect of MRs on review quality. In this context, review quality refers to things like a review's perceived utility in assisting a consumer in making a purchase decision (Mudambi & Schuff, 2010) or the presence of product or service related attributes (Burtch, Hong, Bapna, & Griskevicius, 2018). The quality of online reviews is important for a number of reasons. From the perspective of consumers, it can be frustrating to filter through large

[1]https://tripadvisor.mediaroom.com/2019-12-12-TripAdvisor-Study-Reveals-77-of-Travelers-More-Likely-to-Book-When-Business-Owners-Respond-to-Reviews

quantities of reviews, in order to obtain useful information about the product or service. From the perspective of organizations, high review quality is helpful in order to better understand strengths or weaknesses of their products or services, and in some cases may aid a company in resolving customers' issues or complaints more effectively.

Beyond the scarcity of studies which investigate the effects of MRs on the quality of online reviews, there are two other gaps in the current MR literature, which are addressed in this study. First, the way in which most studies gauge review quality is either by considering the length of reviews (Xu, Li, Law, & Zhang, 2020) or by counting the number of helpful votes reviews receive (Liang & Li, 2019). Even though these measures are valuable and insightful, they fail to capture some of the nuance of what truly makes a review useful to the reader. Second, nearly all of the MR literature stems from the context of the hospitality industry. Specifically, the hotel industry, food & drink services and lodging are dominant industries, for which customer reviews have been analyzed. These fields lend themselves to analyzing customer reviews and MRs, due to the experiential nature of the services which are provided to customers (Sparks & Browning, 2011) and due to the fact that their quality only becomes evident upon consumption (Korfiatis, García-Bariocanal, & Sánchez-Alonso, 2012). However, these things also apply to industries outside of hospitality, thus, an integration of the lessons learned from hospitality literature to other e-commerce contexts is warranted. Therefore, the purpose of this study is to bring about an improved understanding of the effectiveness of MRs in enhancing the quality of online reviews. To this end, the following research question is formulated:

RQ: What is the effect of managerial responses on the quality of online customer reviews?

On the basis of relevant and current strands of literature, including marketing research, information systems research and tourism research, I discuss MRs and online review quality and develop a theoretical framework. The focal MR aspects in this study are the frequency with which companies respond to reviews, the speed with which reviews are answered and the length of the responses. By testing a series of hypotheses, empirical evidence for the effects on four distinct indicators of online review quality - review length, useful votes, readability and diagnosticity - is provided. The results apply eminently to industries centered on internet service providers, such as IT consulting, web hosting, online communications or VPN service providers. However, generalizations can be made about many online spaces of discourse between customers and organizations, which is a particularly relevant contribution to marketing and information systems research.

Several methods are combined in order to retrieve, and analyze a large dataset of online reviews and responses. These include web scraping (R. Mitchell, 2018), qualitative content analysis (Verhoeven, 2016), machine learning (Cielen, Meysman, & Ali, 2016; T. M. Mitchell, 1997) and natural language processing (Sebastiani, 2002), as well as OLS regressions. Reviews are collected from the online review platform Trustpilot. The platform was launched in 2007 and displays customer reviews for a wide range of organizations, spanning from local businesses like restaurants, museums or car dealerships, over specialized online retailers and service providers, to globally operating companies like Uber, AirBnB or Amazon. I elaborate on the platform and its functionality in more detail in the methodology section of this study. To measure review quality, I refer to existing studies on user generated content quality and apply some of their established measures. These are review length, helpfulness and readability. Furthermore, an automatic machine learning classifier is developed, in order to determine whether review texts contain useful product or service related information. Finally, twelve hypotheses are tested using OLS regression models.

1.3 Expected Contributions

I intend to provide several contributions to the current knowledge about MRs and review quality, which are expected to benefit theory and practice. First, previous studies in which MRs have been examined come to different, sometimes contradicting conclusions. It is not universally agreed, for instance, whether responding to a larger or smaller portion of online reviews is advisable. I intend to contribute to clearing up some of these disagreements from the vantage point of the quality of online reviews. The metrics which are commonly used to measure review quality, such as review length and helpful votes, may not always sufficiently capture the true utility of an online review. Therefore, I also consider the presence of service or product related information in review texts.

The second contribution ties into the inclusion of textual review quality characteristics. By employing machine learning and natural language processing techniques, a large dataset of reviews is analyzed. I demonstrate the utility of these approaches, in making sense of large volumes of unstructured text data. Even though machine learning and natural language processing are popular in many research and commercial disciplines (Sheng, Amankwah-Amoah, Wang, & Khan, 2019), thus far, most previous studies related to MRs have not made full use of them.

The third major contribution I intend to provide is the integration of knowledge about MR, which is primarily generated by tourism and hospitality research, into a broader domain

of e-commerce. According to Chen et al. (2019), findings from, for instance, hotel reviews can potentially be generalized to other e-commerce environments. This study aims to test and extend the current knowledge in a context of internet service providers.

Finally, this study contributes to information systems research by shedding light on how firms ought to interact with their customers on third party platforms. The findings of this study are especially relevant for platforms, on which companies do not have full control of the displayed content. Many online spaces of discourse between companies and customers give companies the option of replying to customer comments. This study explores this option and its effects on the quality of online reviews.

There are multiple actors who can benefit from the findings of this study in practical settings. These include businesses which engage in e-commerce, (prospective) customers of these businesses and online review platforms.

Businesses can benefit from specific, high quality feedback about strong or weak points provided by customers. Insights about managerial responses can provide guidance in deciding whether to adopt them as a routine business practice. If the strategy is conducive to review quality, businesses should allocate resources to it accordingly.

For consumers, it can be tedious and inefficient to read many reviews which contain little information about the product or service of interest. Higher quality reviews are better suited to provide information, on which a purchase decision can be based. Additionally, those consumers who take the time to write a review and put effort into it might feel validated and cared for, if their review is acknowledged by a managerial response which might motivate them to continue the discourse with that organization or other organizations (Xu et al., 2020).

Furthermore, review platforms should also have an interest in the effectiveness of MRs in enhancing review quality. If managerial responses do in fact improve review quality, this might bear relevance to the functionality or layout of such review platforms.

1.4 Outline of the Study

In the following section, the theoretical framework for this study is developed. First, I conduct a systematic literature review of MR literature, which is the basis for the independent variable. I continue by briefly reviewing online customer review quality, which is the dependent variable, and defining it for the purposes of this study. Subsequently I develop hypotheses pertaining to MR frequency, speed and length. In section three the methodological strategies and approaches are explained. These include data retrieval via web scraping, the development of an automatic review diagnosticity classifier and OLS regression analysis. In

section four I present the results of the OLS regression analysis. In section five I summarize and interpret the results, provide specific practical advice for companies and review platforms, and discuss theoretical implications, limitations and future research.

2 Theoretical Framework

2.1 Managerial Response: Systematic Literature Review

Some authors have provided literature reviews about managerial responses in the past. One of the most extensive and perhaps influential contributions was made by Davidow (2003). In it, research from as far back as 1982 up to 2001 is reviewed and the author illustrates which dimension of MR is investigated in each article and what the main findings are. I refer to Davidow's (2003) contribution and his six dimensions of service recovery several times throughout this study. A more recent example is provided by van Noort, Willemsen, Kerkhof & Verhoeven (2014), in which the authors focus on webcare for negative customer feedback. The authors discuss whether or not to respond to reviews, when to respond, what to say and what communication style to use. Some of their main findings include that, generally, reacting to customer comments is advisable, but firms should not respond to all comments, since this can be perceived as intrusive. Furthermore Li, Cui & Peng (2017) review some MR literature in the context of their study including several focal variables, effects and main findings for each article. Their review includes 17 articles ranging between 2010 and 2017. To the best of my knowledge, the latter is the most recent review of MR literature. Thus, some additional and extensive review of the current MR literature is warranted.

In Appendix A, I illustrate the key features of all the reviewed articles and summarize the main findings which are pertinent to the domain of MRs[2]. Upon inspecting the Review Context/Industry column of Appendix A, it quickly becomes evident how dominant hospitality literature is in the research on MRs. To be specific, 41 of the 49 articles generate their findings from the hospitality industry with hotels being the most common subject. This lends support to the aforementioned theoretical contribution of this study, to provide insights coming from an e-commerce context other than hospitality. I also illustrate the data sources which are used in each article. These are somewhat diverse, and include a number of scenario based experiments and several online review platforms which are mostly travel related. In

[2] For the methodology of the systematic literature reviews see Appendix B

particular, TripAdvisor and Expedia as well as several Chinese travel platforms, including Qunar, Ctrip and eLong are often used as research contexts. Moreover, I show which particular aspects of managerial responses are investigated in each article and, if applicable, which outcome metrics are observed. In the following sections, I describe and discuss the different aspects of MRs found in the articles.

2.1.1 Aspects related to MR in the Literature

Response Presence vs. Absence

Several studies have investigated contexts in which MRs are provided and compared them to contexts in which no MRs are present. These different settings are elicited in a number of ways. In some studies, scenario based experiments are conducted (S. J. Kim, Wang, Maslowska, & Malthouse, 2016; Rose & Blodgett, 2016; Sparks, So, & Bradley, 2016; Sreejesh, Anusree, & Ponnam, 2019). Other studies compare online platforms on which an MR feature is present to platforms without an MR feature (e.g. Chen et al., 2019; Proserpio & Zervas, 2017). Alternatively, researchers compare responding businesses to non-responding businesses (e.g. Lui et al., 2018) or businesses before and after they started responding to reviews (e.g. Chevalier et al., 2018).

Some of the findings in these studies pertain to review volume and review valence or rating. Chen et al. (2019) find an increased review volume when MRs were provided, which is consistent with findings by Proserpio & Zervas (2017), who report an increase in volume of 12%. Impacts on rating score are not always consistent across the literature. Some researchers report improved rating scores induced by MRs (Proserpio & Zervas, 2017; Sheng et al., 2019), while others do not find significant effects (Chen et al., 2019). According to Sheng et al. (2019) the effect was stronger when responding to negative reviews. Interestingly, Proserpio & Zervas (2017) find that providing MRs overall leads to fewer negative reviews, but the negative reviews which are posted tend to be more detailed. The authors attribute this to a perception of increased scrutiny to the presented criticism, so that reviewers tend to post more carefully argued criticism.

Another area of interest is related to firm performance, for instance financial performance (Kumar, Qiu, & Kumar, 2018) or purchase intentions (S. J. Kim et al., 2016; Sreejesh & Anusree, 2016). According to Kumar et al., MRs lead to increased firm performance of the focal firm and decreased firm performance of other competing local businesses. Sreejesh & Anusree (2016) report increased hotel booking intentions due to MRs, specifically in situations of severe service failure and high agreement among reviewers. Kim

et al. (2016) report increased purchase intentions and decreased negative word-of-mouth intentions for review viewers, but not for review posters.

Lastly, there are studies in which aspects such as perceived trust (Könsgen, Schaarschmidt, Ivens, & Munzel, 2018; Sparks et al., 2016) and company reputation (Rose & Blodgett, 2016) are investigated. Both, Könsgen et al. (2018) and Sparks et al. (2016), report an increase in perceived trust induced by MRs. The former study investigates a context of employee reviews about their employer and also finds increased intentions to pursue employment, induced by MRs. In the latter study, a hotel context is investigated and the authors further report an increase in perceived concern for customers (attentiveness, caring & responsiveness). According to Rose & Blodgett (2016), MRs to negative reviews are effective in improving company reputation, specifically when problems described in customer reviews are perceived to be controllable.

One of the most important strands of literature, on which a number of these studies and studies about management responses in general draw, is service recovery literature. Davidow (2003) made an impactful contribution to the field, proposing six dimensions of organizational responses to complaints; timeliness, facilitation, redress, apology, credibility and attentiveness, which influence customer word-of-mouth behavior. I will come back to some of these dimensions throughout the following paragraphs, briefly explaining their relevance to the respectively discussed aspect. The dimension 'facilitation' is the one which is most applicable here. It refers to a company's policies, procedures and structures to support customers engaging in complaints and communications. Whether or not companies engage in communication with customers on third party platforms can be seen as a procedure related to this dimension. Facilitation is considered to be conducive to favorable word-of-mouth behavior. Many of the studies reviewed above design their frameworks or hypotheses about the use of management responses in reference to Davidow (2003) and, more or less explicitly, the 'facilitation' dimension.

Response Frequency

Another fairly common aspect is the frequency with which responses to reviews are posted. This is often investigated in a context of online review platforms, with TripAdvisor being by far the most popular one (e.g. Alrawadieh & Dincer, 2019; Lee, Besharat, Xie, & Tan, 2018; Schuckert, Liang, Law, & Sun, 2019; K. Xie, Kwok, & Wang, 2017; K. L. Xie, Zhang, Zhang, Singh, & Lee, 2016).

Once again, effects on review volume and review rating are investigated by several authors. All researchers who investigate review volume report increased volume when MR frequency is higher (Chevalier et al., 2018; C. Li et al., 2017; Sheng, 2019; K. L. Xie et al., 2016) and most researchers who investigate rating find that higher MR frequency leads to higher ratings (C. Li et al., 2017; Liang, Schuckert, & Law, 2017; Schuckert et al., 2019; K. L. Xie et al., 2016). However, one study also finds the opposite effect on rating scores (Chevalier et al., 2018), suggesting that negative consequences can occur in certain circumstances. The authors of this particular study mention that many businesses respond predominantly to negative reviews, which may encourage customers who had negative experiences to voice their complaints. Contrary to this finding, some researchers recommend responding to negative reviews more frequently, in order to achieve stronger improvements in rating scores (e.g. Schuckert et al., 2019)

Outcomes pertaining to firm performance are not entirely uniform in the literature either. While some researchers report increased revenue or bookings due to MRs (Lui et al., 2018; K. Xie et al., 2017; K. L. Xie et al., 2016) others do not find significant effects (Z. Zhang, Li, Meng, & Li, 2019) or even find negative effects on revenue (Lee et al., 2018; Xu, Zhang, Law, & Zhang, 2019). According to Xu et al. (2019), the decline in firm performance can be attributed to the fact that there is no purchase verification mechanism on the platform they investigated, causing distrust among review readers. Lee et al. (2018) specify that MRs tend to decrease revenue if review rating and volume are both low or both high. In situations where one is high and the other is low, higher frequency tends to increase revenue.

The last outcome measure I draw attention to is reviewing effort, since this is highly relevant for this study. Two studies present results which pertain to the effect of MR frequency on reviewers' effort put into writing reviews (Chevalier et al., 2018; Xu et al., 2020). Both of them measure reviewing effort using the length of the review. Xu et al. (2020) additionally look at the 'expert review' attribute, which is a feature on the travel review platform they investigate (qunar.com). Both studies find increased reviewing effort caused by higher MR frequency. According to Xu et al., (2020) the effect can be attributed to users' activated sense of reciprocity when they receive an MR. In order to give back, they tend to provide higher quality reviews to the platform and the readers. It is noteworthy that these authors investigated users' profiles and the effect MR frequency had on users' general reviewing behavior on the platform. So, for instance, receiving an MR from one hotel might affect that user's future reviewing effort for other hotels on the platform. According to

Chevalier et al. (2018), a main driver for the increase in review length is the motivation to reach managers and improve quality of the service in this way.

Response Speed

Seven studies in the set of articles measure effects of response speed (Alrawadieh & Dincer, 2019; C. Li et al., 2017; Sheng, 2019; Sheng et al., 2019) where the date of the posting of the customer review and the date of the response are recorded and the span between them marks the response speed. More studies use datasets with information on response speed but they don't necessarily focus on the effect of it, however, some use it as a control variable (e.g. Lui et al., 2018; L. Zhang, Gao, & Zheng, 2020).

Response speed has been regarded as an indicator of an organization's efficiency in previous literature (Sparks et al., 2016). Furthermore, service recovery theory (Wallin Andreassen, 2000) suggests that, in situations of service failure, the speed with which such failures are handled is important. Another relevant theory in this context is the theory of media synchronicity (Dennis, Fuller, & Valacich, 2008) where shared, coordinated behavior leads to better communication performance, i.e. a shared interpersonal understanding for all parties involved. One of the theory's aspects is transmission velocity, which refers to the speed at which messages are delivered and replied to. Higher transmission velocity is said to improve communication performance, as it more closely resembles conversation. When responding to online reviews, the theory would advocate quick responses to achieve better communication performance. Coming back to Davidow's (2003) six dimensions, 'timeliness' is the dimension which is relevant here. It is the perceived speed with which a company responds to or handles a complaint. Timelier responses are said to be conducive to more successful service recovery. According to Sheng (2019) and Li et al. (2017), higher response speed increases review volume. When it comes to rating score, there are two studies which report diverging results. Sheng et al. (2019) did not find response speed to significantly affect rating score, whereas Li et al. (2017) report an increase in rating score for higher response speed. Financial performance has been investigated in two separate studies, both of which report increases in revenue with higher response speed (Sparks et al., 2016; K. Xie et al., 2017). Finally, Sparks et al. (2016) find that higher response speed increases customers' perception of attentiveness, caring and responsiveness, which induce trust in the company. Specifically, they find that the most positive outcomes are achieved when the time lag is within one day, as opposed to one week or one month.

Response Style

A decent portion of the reviewed articles investigates different response styles found in MRs and their effects on different metrics. Some studies explore which response styles are used by managers, using qualitative research designs (e.g. Alrawadieh & Dincer, 2019; Feng & Ren, 2019; Ho, 2017; Sparks & Bradley, 2017) and building frameworks based on their findings. Others compare the effectiveness of different response styles on various outcome variables and in various situations. Most authors use scenario based experiments or questionnaires to achieve this (e.g. Marx & Nimmermann, 2018; Olson & Ro, 2019; Piehler, Schade, Hanisch, & Burmann, 2019; W. Weitzl, Hutzinger, & Einwiller, 2018; W. J. Weitzl, 2019), while some use reviews and MRs retrieved from online review platforms (e.g. Casado-Díaz, Andreu, Beckmann, & Miller, 2020; Lui et al., 2018). Many of these studies draw on the aforementioned service recovery literature, specifically by Davidow (2003), in order to propose their frameworks, hypotheses or selection of response strategies.

Among the qualitative exploratory studies, some pay special attention to negative online reviews (e.g. Alrawadieh & Dincer, 2019; Mate, Trupp, & Pratt, 2019; Sparks & Bradley, 2017). Alrawadieh & Dincer (2019) identify four main components of responses to negative reviews; gratitude for the feedback, apology, explanation and some form of incentive or compensation. They find that gratitude and apology are the most common components, followed by explanation. Sparks & Bradley's (2017) typology entails several forms of acknowledgement of the negative event, several types of explanation of the event and several types of actions following the event. Some of the most common response components within this typology include expressing gratitude for the feedback, recognizing the event, apology, justification, denial and excuse. In a study differentiating between most common response types to negative vs. positive reviews, Feng and Ren (2019) find that thanking, advertisement and promising responses were most common for positive reviews whereas justification and offer of solution were most common for negative reviews.

Some of the most commonly tested response strategies are 'accommodative', 'defensive' & 'no response' strategies. Several studies test one (e.g. Piehler et al., 2019; Rose & Blodgett, 2016) or more (e.g. Casado-Díaz et al., 2020; Li, Cui, & Peng, 2018; Meng, Dipietro, Gerdes, Kline, & Avant, 2018; Weitzl, 2019; Weitzl & Einwiller, 2019) of these strategies' effects on certain metrics. Some authors find accommodative strategies to be most effective in mitigating negative word-of-mouth intentions or brand satisfaction, particularly if complaints have a constructive tone (Weitzl, 2019), if reviewers are constructive and loyal customers (Weitzl & Einwiller, 2019) or if few prior service failures have occurred (Weitzl et

al., 2018). These three studies are consistent with Li et al. (2018) and Casado-Díaz et al. (2020), which report accommodative responses to be most effective to improve hotel revenue and purchase intention (C. Li et al., 2018) and brand attitude (Casado-Díaz et al., 2020). In cases where negative reviews do not refer to product or service features, but are vindictive or unconstructive, no difference in effects of response style was found (Weitzl, 2019; Weitzl & Einwiller, 2019). Finally, Piehler et al. (2019) focus exclusively on accommodative response strategies and differentiate between explanation and compensation. They find that both have a positive impact on purchase intention of potential customers, i.e. observers of reviews, rather than review writers. They also report that a combination of the two has the strongest positive impact. The impact on review viewers is attributed to signaling theory, according to which potential customers search for credible signals when information about a service or product is not obvious prior to consumption.

The related dimensions in Davidow's (2003) framework are redress, apology and credibility, all of which fit the parameters of accommodative response strategies. Credibility refers to an organization's willingness to present an explanation for the problem. Redress refers to the benefits or response outcomes a customer receives from the organization and apology is defined as an acknowledgement of the complainant's distress (Davidow, 2003).

Response Length

Seven of the selected articles research MR length (C. Li et al., 2017; X. Liu & Law, 2019; Schuckert et al., 2019; Sheng, 2019; Sheng et al., 2019; K. Xie et al., 2017; K. L. Xie, So, & Wang, 2017). All of these studies examine reviews and MRs retrieved from travel review platforms in the context of the hotel industry. While one study employs an exploratory approach, in order to find out how much effort managers exert towards different types of customer reviews (Liu & Law, 2019), the others measure effects of MR length on review volume (C. Li et al., 2017; Sheng, 2019), rating (C. Li et al., 2017; Schuckert et al., 2019; Sheng et al., 2019) and financial performance (K. Xie et al., 2017; K. L. Xie et al., 2017).

According to Liu & Law (2019), managers exert more effort, as measured by MR length, towards negative reviews. To be specific, a one-score decrease leads to a 17-word increase in managerial responses. Moreover, the authors find that managers exert more effort when responding to longer reviews.

The two studies which measured effects on review volume find diverging results. While Sheng (2019) reports that higher MR length leads to higher review volume, Li et al., (2017) did not find higher MR length to significantly increase review volume. For review

rating, Sheng et al. (2019) and Schuckert et al. (2019) both find a positive effect when MRs were longer. The latter study specifies that the effect applies to MRs to negative reviews in particular. Finally, the results pertaining to financial performance are not uniform. According to Xie, So & Wang (2017), lengthier MRs increase revenue, whereas Xie, Kwok & Wang (2017) find higher MR length to decrease revenue. So it can be said that effects of MR length on various metrics are not conclusively researched.

One theory which provides grounds for expected effects of longer MRs is uncertainty reduction theory. Higher length of MRs are expected to go along with more informational content, which in turn increases the message's capacity to reduce uncertainty for the parties involved in the communication process (Daft & Lengel, 1984). For customers who observe interactions between the firm and other customers, a high level of detail in the MR might signal that the firm cares strongly about its customers and their experiences. This perception might incite customers' engagement and motivate them to share their own experiences (Li et al., 2017).

Personalization

The last aspect, which is discussed here pertains to personalization of MRs to the reviews and to what extent MRs are specific or standardized. In some studies, this aspect is closely associated with an MR's capacity to inform and reduce uncertainty (Xie et al., 2017). Zhang et al. (2020) investigate effects on rating score and find that higher levels of matching responses increase rating. Two studies investigate effects on customers' intention to co-create, using experimental designs (Roozen & Raedts, 2018; Shin, Perdue, & Pandelaere, 2019). They both report personalized MRs to be more conducive than highly standardized responses. According to Roozen and Raedts (2018), a personalized response is particularly important when answering to reviews with a mix of positive and negative evaluations, whereas purely negative or purely positive reviews benefit less from highly personalized MRs. Effects on firm performance have been measured in four studies. Zhang et al. (2020) frame the issue in terms of similarity between MRs and find that high similarity in MRs has a negative effect on hotel bookings. Li, Cui & Peng (2018) lend support to these findings, reporting that more tailored MRs improve purchase intentions and hotel revenue. However, there are also authors which report deviating findings, for instance, Xie, Kwok & Wang (2017) find no significant effect of match rate between review and MR on financial performance. In another study (Xie et al., 2017) the authors even report potential negative effects on financial performance, for higher repetition of topics between reviews and MRs. The authors note that it is imperative to

offer additional constructive points, rather than merely repeating issues raised in the review. Finally, Raju (2019) investigates the effects on review readers' perceived fairness when faced with vague MRs vs. specific MRs. He finds that specific MR content leads to higher perceived fairness than vague MR content.

Theoretical justification for the results can be drawn from the elaboration likelihood model (Petty & Cacioppo, 1986). Specific and personalized MR content can be considered as a central cue, as it is expected to be relevant to the receiver of the message. Thus, the receiver's (or reader's) motivation can be affected by specific or personalized content of MRs.

2.2 Online Customer Review Quality: Previous Work and Concepts

The quality of user generated textual content in online platforms is a subject of attention in several strands of literature such as marketing science, information systems research and management science. In this section, I discuss several concepts associated with quality, commonly found in these research contexts. Bearing in mind the focus of this study, namely online reviews and MRs, I describe four relevant components of content quality, which are present in previous studies. By considering multiple concepts related to quality, I ensure a nuanced view of review quality. Specifically, the four concepts I discuss are review length, helpfulness, diagnosticity and readability.

2.2.1 Review Length

The length of user generated textual content is a common concept associated with quality. It is widely held that longer user generated customer reviews correspond to higher quality (e.g. Burtch et al., 2018; Mudambi & Schuff, 2010; Pan & Zhang, 2011). According to Goh, Heng & Lin (2013), longer reviews are more likely to contain considerable product information. In practice, however, customer reviews are often very brief, limiting their utility for other consumers (Cao, Duan, & Gan, 2011; Mudambi & Schuff, 2010). Mudambi & Schuff (2010) further elaborate that the added utility of longer, more detailed reviews lies in enabling readers a more confident purchase decision and a reduced product/service quality uncertainty. Furthermore, longer reviews may indicate that the review writer has exerted more effort towards writing the review and is more involved in providing high quality information to aid other people (Pan & Zhang, 2011). Some studies have investigated the effect of different strategies to motivate users to write longer reviews. Two such studies report that financial incentives can lead to shorter reviews whereas socially motivated strategies are better suited to induce lengthier reviews (Burtch et al., 2018; Khern-am-nuai, Kannan, &

Ghasemkhani, 2018). Specifically, socially motivated strategies of those two studies include introducing larger audiences, audiences consisting of peers and emphasizing social norms in writing reviews. Tversky and Kahneman (1974) suggest that the availability of more information about or reasons for a decision will enhance the confidence of the decision maker. Extrapolated to online reviews, this suggests longer reviews to be more conducive to a confident purchase decision. Finally, Petty & Cacioppo (1984) posit that a message that is processed with a higher cognitive intensity is more likely to lead to an attitude change. Accordingly, a longer online customer review might require more cognitive processing and might in turn be more convincing than a short review.

2.2.2 Review Helpfulness and Diagnosticity

The term helpfulness is often referred to in relation to online reviews in practice and research. Online retailers like Amazon or third party review websites like TripAdvisor or IMDB commonly use the term when asking readers to evaluate reviews, although there is not necessarily an explicit uniform definition of the term across platforms. In information systems and marketing research, it is described as the extent to which a review facilitates the consumer's purchase decision process (Mudambi & Schuff, 2010) or the extent to which a consumer perceives a review to be useful in performing his/her shopping task (Pan & Zhang, 2011). While some studies use consumer perception as the sole indicator for helpfulness (Forman, Ghose, & Wiesenfeld, 2008), others have also considered structural, lexical or semantic factors of the review texts (Kim, Pantel, Chklovski, & Pennacchiotti, 2006). Despite the differences in how helpfulness is measured across the aforementioned studies, they uniformly consider higher helpfulness to be conducive to review quality. Bearing in mind that I consider text related characteristics in connection with the other three concepts, I choose to limit helpfulness to the consumer perception aspect for this study.

A concept which is closely associated with helpfulness in information systems and marketing literature is related to whether a review contains product or service related characteristics. Several studies refer to this as diagnosticity (Burtch et al., 2018; Mudambi & Schuff, 2010). The difference to the previously discussed helpfulness concept is subtle but important. Whereas helpfulness refers to the extent to which a review facilitates the purchase decision, diagnosticity specifically emphasizes whether a review informs about product or service related attributes (Burtch et al., 2018). Depending on the product or service category, this may include descriptions and opinions about price, quality, utility, aesthetics, customer support, delivery or setup speed and staff performance to name just a few examples. This is

important because it enables consumers to evaluate a product's or service's performance prior to purchasing it. Thus, diagnosticity is a bit more objective compared to a consumers' evaluation of perceived helpfulness in making a purchase decision. While some studies consider diagnosticity to be part of a review's helpfulness (Mudambi & Schuff, 2010) others make a clear distinction between the two concepts (Burtch et al., 2018). However, diagnosticity is consistently related to qualitative review text features in all of the studies mentioned here. It should also be noted that not all studies explicitly refer to the term diagnosticity when discussing whether reviews describe service or product related attributes. For instance, Fan and Li (2006), in developing a machine learning classification approach to identify low quality reviews, refer to a review's presentation of product features as one dimension by which it can be considered of high quality. Similarly, Goh et al. (2013) refer to content information richness as the amount of information about product or brand attributes and usage experience. These descriptions closely resemble the definition of diagnosticity presented above. Revisiting considerations of uncertainty reduction (Tversky & Kahneman, 1974), high diagnosticity in reviews should be effective in mitigating uncertainty, as it allows consumers to access product and service related information. Based on the discussion of the differences between the two concepts, a distinction is made between helpfulness and diagnosticity for the purposes of this study

2.2.3 Review Readability

Finally, readability is a concept that has been considered a reliable indicator of review quality in previous studies (e.g. Khern-am-nuai et al., 2018; Korfiatis et al., 2012). Readability refers to the ease of understanding or comprehending a text and describes the effort and educational level required from the reader (DuBay, 2004; Korfiatis et al., 2012). There are a number of readability formulas which estimate, for instance, the amount of years of education required to understand a given text, based on its linguistic and structural characteristics. Specifically, characteristics like word complexity, number of words or number of sentences are key aspects in many of these formulas. Beyond these aspects which are inherent to the text itself, some modern views of readability steer the focus more towards characteristics of the reader (Pikulski, 2002). For the purposes of this study, however, the ability to predict review quality provided by readability formulas is adequate. According to Korfiatis et al. (2012) readability formulas should preferably be used to evaluate short texts, making them ideal in the context of online customer reviews. Several studies have used readability tests to assess online reviews (e.g. Khern-am-nuai et al., 2018; Korfiatis,

Rodríguez, & Sicilia, 2008; Korfiatis et al., 2012; Liu et al., 2007), unanimously considering reviews with high readability more reliable or credible. Other authors mention positive effects of higher online review readability on sales of digital cameras (Ghose & Ipeirotis, 2011) and on the number of helpfulness votes (Forman et al., 2008).

2.3 Hypothesis Development

On the basis of the preceding discussion about online review quality and the extensive review of recent MR literature, I now develop hypotheses pertaining to MR frequency, MR speed and MR length and their effects on the quality of online reviews. Note that the theorized effects mainly refer to the impact of MRs on the quality of other people's reviews, rather than the review which received the response. That is, by observing MRs to reviews of other users, subsequent review quality is affected. The framework and hypotheses are visually displayed in Figure 1.

2.3.1 Response Frequency

Among the reviewed literature, there are two studies in which effects of MR frequency on review length are presented (Chevalier et al., 2018; Xu et al., 2020). Both of them find MR frequency to lead to increased review length. The presence of MRs to customer reviews signals to the reader that the organization reads the reviews and is interested in the feedback provided by customers (Chevalier et al., 2018; Lui et al., 2018). In the eyes of a potential review writer, this may lead them to believe that writing a review is more likely to have an impact on things like future performance or behavior of the organization or even their own interaction with the organization. This may prompt them to not only write a review, but also pay close attention to the quality of their review. A higher frequency of MRs may increase the perceived likelihood that posting a review will have real effects on the company and perhaps be met with a response.

Following findings of a previous study (Proserpio & Zervas, 2017), I also argue that reviewers who intend to voice complaints in their review will expect increased scrutiny towards criticism, if they see high MR frequency. Consequently, they will tend to post more carefully argued criticism, in order to call attention to shortcomings or to get their issues resolved. This is consistent with findings from service recovery literature, that the ease with which customers can resolve complaints is important for customers' word-of-mouth behavior (Davidow, 2003). In this context, high frequency of MRs signals a company's willingness to acknowledge and resolve complaints. Apart from effects on length, as observed in previous

studies, I also expect positive effects on the other review quality characteristics. Thus, I hypothesize that review writers will write longer, more readable, more helpful and more diagnostic reviews if MR frequency on a business profile is higher:

Hypothesis 1: Higher MR frequency increases the quality of customer reviews.

(1a) Higher MR frequency increases the length of customer reviews.

(1b) Higher MR frequency increases the helpfulness of customer reviews.

(1c) Higher MR frequency increases the readability of customer reviews.

(1d) Higher MR frequency increases the diagnosticity of customer reviews.

2.3.2 Response Speed

The literature reviewed earlier tends to report favorable outcomes for higher response speed on metrics such as financial performance, review rating, perceived firm attentiveness and review volume. It has been argued that higher response speed symbolizes a firm's active embrace of managing customer comments (Sheng, 2019). To review readers, quick replies indicate that an organization tries to maintain an interactive relationship with customers. This has been referred to as positive inferences about an organization's concern for its customers (Sparks et al., 2016). It is reasonable to expect that reviewers are more likely to provide high quality reviews, if they perceive a company to be responsive and committed to their customers in this way. This argument is related to the aforementioned theory of media synchronicity, where higher transmission velocity (i.e. faster response) contributes to an improved understanding for all parties involved (communication performance), because it more closely resembles human conversation (Dennis et al., 2008). Further support can be drawn from the service recovery literature. Davidow (2003) posits that long delays in responses may be detrimental to consumers' subsequent word of mouth behavior, even though acceptable response times may be context specific.

Another argument is the visibility of a review and its following response on the review platform (De Vries, Gensler, & Leeflang, 2012; Sheng, 2019). This is especially pertinent for the review platform Trustpilot, since it always displays the twenty most recent reviews on the first page of each company profile. With multiple reviews being posted each day and reviews being displayed in reverse chronological order (most recent reviews appear first on each page), a review will most likely only be displayed on the first page of a company profile for a brief period. Quicker responses will increase the visibility of the interaction between the reviewer and the firm, since customers tend to stay within the first few pages (Pavlou & Dimoka, 2006; Sheng, 2019). If customers notice that a company replies more quickly to

reviews, this may lead them to believe that their interaction with the company is more visible to others. Thus, reviewers may be prompted to pay closer attention to the quality of their review:

Hypothesis 2: Higher MR speed increases the quality of customer reviews.

(2a) Higher MR speed increases the length of customer reviews.

(2b) Higher MR speed increases the helpfulness of customer reviews.

(2c) Higher MR speed increases the readability of customer reviews.

(2d) Higher MR speed increases the diagnosticity of customer reviews.

2.3.3 Response Length

The reviewed studies, which investigate the impact of MR length on metrics like review volume, review rating and financial performance, find somewhat diverging effects. Results tend to suggest that longer reviews are more favorable, although a few authors report inconclusive or even opposite results. Uncertainty reduction theory has been referred to by some of these studies to explain the observed effects (C. Li et al., 2017). It posits that messages with more information richness are better suited to reduce uncertainty and ambiguity for the receiver of the message (Daft & Lengel, 1984). In the context of online reviews, receivers also include users, who observe the interaction between the company and the reviewer. Longer messages are more likely to be richer in information and, thus, have a higher potential to reduce uncertainty. If (potential) reviewers observe longer, information rich MRs to other peoples' reviews, they might be inclined to share their own experiences and feedback with more detail and quality, expecting to receive longer, information rich MRs themselves. That is, in hopes of receiving detailed, insightful responses, users tend to post more detailed, helpful, well written and diagnostic reviews.

The previous argument focuses on the externality of customer reviews and MRs, meaning that an MR to one review impacts other, future reviews. Additionally, there is an argument to be made that MR length may impact review quality of the focal review directly. This has to do with the functionality of Trustpilot. Review writers can edit their own review at any point. In this way, reviewers can, for instance, react to the points made in the MR or give an update to provide information on how an issue was handled by the company. Trustpilot recommends that, if users make updates to their review, they should keep the original review as it was and mark any additions as updates. Thus, others can conveniently observe the

interaction between the customer and the organization[3]. It should be noted that users are in no way obliged to adhere to this recommendation, every user who wishes to edit his or her review is free to do so however he or she chooses. Nonetheless, it is reasonable to assume that longer MRs provide more incentive for reviewers to edit their review and reply to the points made in the MR. In addition to the externality effect, this would presumably make reviews longer and more diagnostic. Moreover, it may make reviews more readable since the reviewer is now in communication with a manager and might formulate their reply more deliberately. Based on these arguments, I expect longer MRs to positively impact review quality:

Hypothesis 3: Higher MR length increases the quality of customer reviews

(3a) Higher MR length increases the length of customer reviews.

(3b) Higher MR length increases the helpfulness of customer reviews.

(3c) Higher MR length increases the readability of customer reviews.

(3d) Higher MR length increases the diagnosticity of customer reviews.

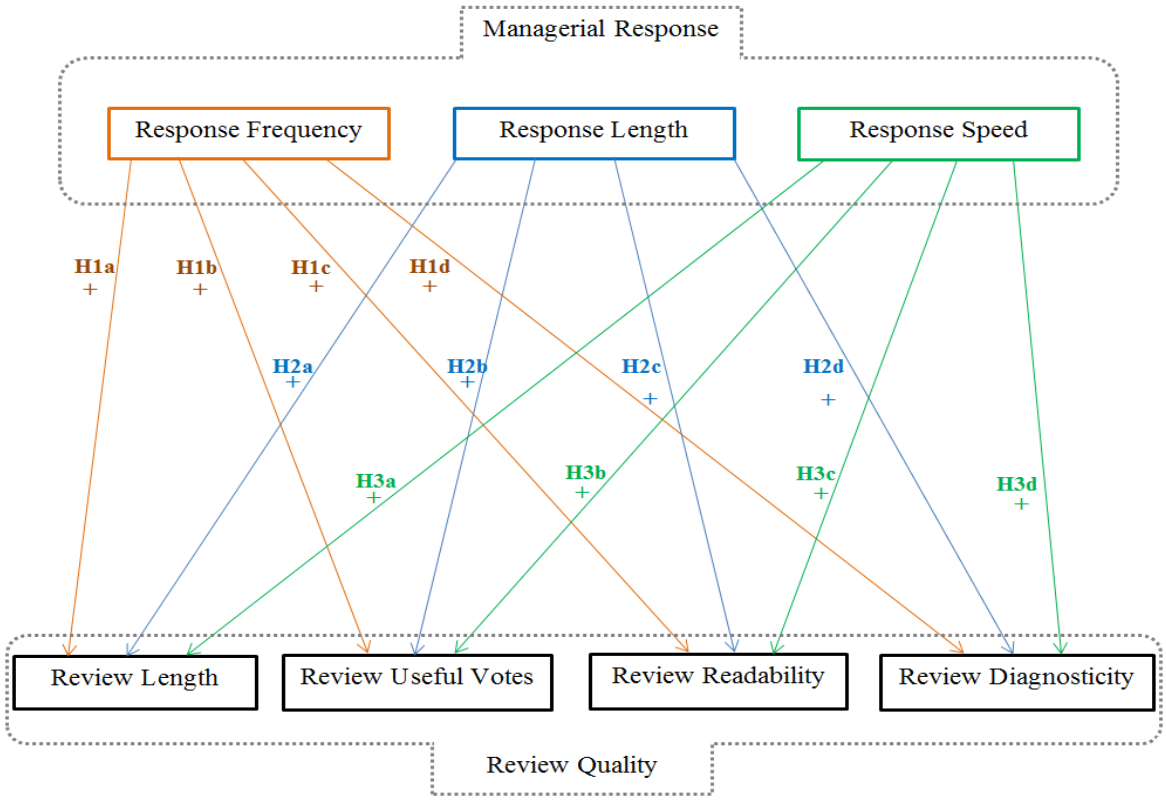

Figure 1 Model Framework

[3]https://support.trustpilot.com/hc/en-us/articles/115015645148-Can-companies-respond-to-reviews-#conversation-1

3 Research Design & Methodology

In this study, multiple methods are combined to analyze the effect of MRs on the quality of online reviews. In the following sections, I elaborate on each of these methods. First, in order to retrieve online customer reviews and responses, I make use of web scraping. In this step, I extract several data points from each online review, including review and response texts and compute a few characteristics, for instance, the review length and the response speed. I also explain in some detail how the readability scores are calculated. Second, in order to determine the diagnosticity of large quantities of online reviews, I develop an automatic machine learning text classifier, by combining qualitative content analysis with natural language processing. Third, several OLS regressions are conducted, in order to test the hypotheses developed in the previous section and estimate effects of MRs on the four online review quality metrics introduced in section 2.2.

3.1 Research Context: Trustpilot.com

The research context of this study is the independent review platform Trustpilot.com. Trustpilot is a consumer review website which hosts reviews for businesses worldwide ranging from local businesses like restaurants, museums or car dealerships, over specialized online retailers and service providers, to globally operating companies like Uber, AirBnB or Amazon, making all reviews openly viewable. The platform was launched in 2007 and displays consumer reviews for over 320,000 businesses and organizations[4]. This abundance of companies and the diversity in terms of industries make Trustpilot a great resource to study online reviews. Another key argument for choosing Trustpilot as the research context is the presence of a response feature, which companies are free to utilize however they see fit. Furthermore, the platform makes certain company and reviewer characteristics visible, which are relevant for the present study and not necessarily apparent on other reviewing platforms.

All of the online reviews for a specific company are displayed on that company's Trustpilot profile. In principle, every user can create a business account and set up a company profile, even if the business is not their own. But the profile also indicates whether the company is 'Unclaimed', 'Claimed' or 'Asking for reviews'. Unclaimed profiles are not managed and maintained by the companies or organizations because the company might not be aware of the profile or just has not claimed it yet. So for these profiles, there is no way of knowing whether the company is actually affiliated with the profile. In order to receive the

[4] https://www.trustpilot.com/about

'Claimed' status, Trustpilot has to verify that the user who runs the profile is in fact the owner of the business. Thus, owners or managers can claim and acquire the profile of their organization, even if they did not create the profile themselves. Claimed profiles are managed by the company or organization, but they do not actively invite customers to review them on Trustpilot. 'Asking for reviews' means that the company or organization manages and maintains the profile and invites their customers to review them on Trustpilot[5]. For this study, only profiles of the two latter categories are considered, since only on these profiles managerial responses to customer reviews can be found.

In order to be posted, reviews must be at least eleven characters long. In addition to writing the review, users must also rate the company with one to five stars. So by default, some effort, albeit rather small, is required before a review can be posted. Other discernible features of the reviews, as shown in Figure 2, include the date on which the review was posted, the number of useful votes it received, the total number of reviews a user has posted on Trustpilot, whether the review is a verified purchase and the number of reviews a user has posted on Trustpilot. Finally, right below the review, companies can post a response, for which the date is indicated. The responses are limited to one response and reviewers cannot post additional answers under the company's response. They can, however, edit their original review and react to the points made in the managerial response. Just like the customer reviews, all managerial responses are openly displayed on the company profile.

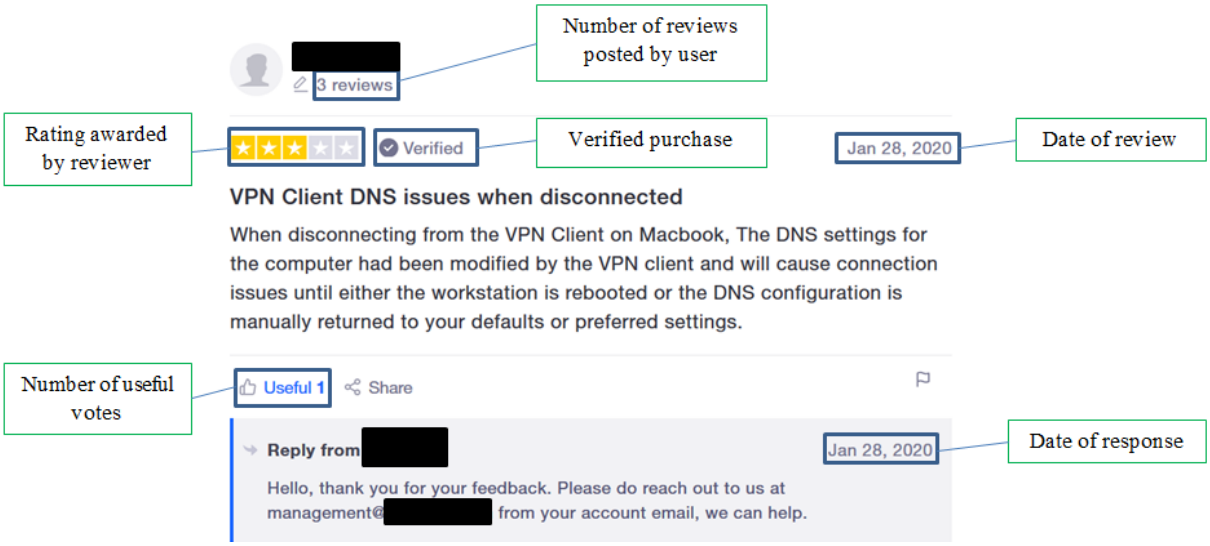

Figure 2 Example of Trustpilot Review & Response

[5]https://support.trustpilot.com/hc/en-us/articles/219386577-What-do-Asking-for-reviews-Claimed-and-Unclaimed-mean-

3.2 Review Collection & Data Structure

In order to obtain customer review data from Trustpilot, I designed a web scraper (R. Mitchell, 2018) using several packages and libraries from the python programming language[6], including "BeautifulSoup" and "Requests". The scraper accesses and parses HTML text to retrieve several items from every consumer review of any given Trustpilot company profile. It retrieves the review texts, posting dates, rating scores and the number of useful votes. Finally, if a response is present, it retrieves the response text and response posting date. These data points allow me to compute review and response word counts, readability scores and, if applicable, number of days between review and reply. In order to obtain some additional control variables, I retrieved the status of each company profile, whether a review is verified and the total number of reviews a reviewer has posted on Trustpilot[7].

Based on predetermined selection criteria, I specified a list of company profiles, from which I scrape reviews and the above mentioned data points. First, I selected a relevant business category. Trustpilot features over four thousand categories and sub-categories, which list comparable businesses[8]. For the purposes of this study, categories related to intangible service goods are most interesting, since reviews are especially useful in these contexts as their utility only becomes evident upon consumption (Korfiatis et al., 2012). This is also in line with the theme of extrapolating findings from the hospitality literature to an e-commerce context, since hospitality is a strongly service driven field. A category which nicely fits this idea is the category 'Internet Service Provider'. Companies in this category offer a range of online services, including VPN services, online communication services, web hosting services, mobile network services and online software services. Thus, the business category 'Internet Service Provider' was selected.

The next selection criterion is that the company profile must be 'claimed' or 'asking for reviews', since only these profiles can have managerial responses and are definitively affiliated with the actual company.

Furthermore the number of reviews on the company profile must be over 500. This ensures that profiles of companies with very few reviews do not enter the data set. I decided to specify this limitation because it allows for viewers of company profiles to observe the way reviewers and firms interact. If a profile only has very few reviews, there is little chance for

[6] My python codes can be accessed via https://github.com/ProfessorDonLuigi/ScrapeAndClassify
[7] There were seven instances in the dataset where the total number of user reviews was 0, which is not a logical value. These turned out to be mistakes in the Trustpilot html text. I corrected these seven values manually.
[8] https://support.trustpilot.com/hc/en-us/articles/360022026634--Categories-and-filters

consumers to observe interactions. Since a key aspect I intend to investigate is related to how observing these interactions influences reviewers in writing their own review, it makes sense to ensure an ample number of reviews. Moreover, it is useful to have a sufficiently high number of reviews representing each company, so that all companies can affect results somewhat equally. It should be noted that this number refers to reviews written in English, not overall reviews. Trustpilot collects reviews in many languages, however, in this study only reviews written in English are collected, which is a filtering option provided by the platform.

Lastly, I omitted all reviews from the first page of each company profile, because many of them are posted hours or even minutes before the moment of data collection. For these very recent reviews companies might not yet have had the chance to post a response. This selection and collection process resulted in 170,340 consumer reviews and 26,508 responses from 55 company profiles. All the review data was anonymized so that no company or user names are apparent in this research.

3.3 Readability Formula Consensus

Previous studies have established that higher readability scores, which correspond to more complex texts, are associated with higher review quality (Khern-am-nuai et al. 2018; Korfiatis et al., 2008).

In order to measure the readability of consumer reviews, I use a consensus of eight usability formulas, which estimates the amount of years of education required to understand a text upon reading it once. This readability consensus is a function provided by the python package "Textstat", which combines the Flesch-Kincaid grade level (Kincaid, Fishburne, Rogers, & Chissom, 1975), Gunning-Fog index (Gunning, 1952), Flesch reading ease formula (Flesch, 1979), SMOG index (McLaughlin, 1969), automated readability index (Smith & Senter, 1967), Coleman-Liau index (Coleman & Liau, 1975), Linsear Write formula (Klare, 1974) and Dale-Chall readability (Dale & Chall, 1948) score. Previous studies have used combinations of readability formulas (e.g. Gyasi, 2013). Burke and Greenberg (2010) advocate the use of multiple readability formulas because some formulas weigh certain text characteristics more heavily than others, leading to potentially large differences between formulas for a given text. The Gunning-Fog index, for instance, puts a lot of weight on the number of complex words, with complex words being words with more than two syllables. Even though this index has been very popular in previous studies on online reviews (e.g. Goes, Lin, & Yeung, 2014; Khern-am-nuai et al., 2018) due to its ease of use (DuBay, 2004), it tends to overrate certain types of texts. Very short texts with one or two complex words

may receive exaggerated readability scores, which can be problematic for the dataset used in this study. Taken by themselves, each of these formulas may have certain weaknesses, however, combining them can help provide a more balanced picture. Most of the formulas take into account the average number of syllables per word, the number of words in the text and the average number of words in each sentence (DuBay, 2004). Alternatively, some tests, like the Dale-Chall readability score, consider the inclusion of certain words from a list to assess word complexity (Dale & Chall, 1948).

3.4 Classifying Diagnosticity of Online Reviews

In this section, I elaborate on how I determine the diagnosticity of online customer reviews by combining qualitative content analysis with machine learning and natural language processing (Priante et al., 2016).

3.4.1 Qualitative Content Analysis & Manual Annotation

One of the early steps in the development of the online review diagnosticity classifier is to precisely define the parameters of what constitutes a diagnostic review. One study in which online reviews have been classified in a comparable way is provided by Liu et al. (2007), who developed a framework to identify low quality reviews, involving machine learning classification. They do not explicitly refer to the term diagnosticity, however, one of the aspects by which they determine review quality is related to product specific aspects and opinions present (or absent) in a review. Essentially, their minimum requirement for a review to be considered of high quality is that it provides a very brief description of the product and comments on or evaluates some aspects of the product. This is a good starting point for the purposes of this study. However, since the dataset in this study exclusively consists of customer reviews for services, I make some adjustments to the parameters specified by Liu et al. (2007). I consider whether a review includes comments, descriptions or opinions about any aspects related to the service or experience with the company. Something akin to the aforementioned ‘very brief description of the product’, as proposed by Liu et al. (2007), is not always applicable in the context of Trustpilot reviews. This is because the reviews are sometimes written about the company, not necessarily about individual products or service goods. Furthermore, many Trustpilot company profiles feature brief descriptions of the company and the kinds of products or services they offer, making a brief company or service description redundant in some cases. Thus, I consider a brief description of the company or service to be conducive but not required, in order for a review to be of high diagnosticity. If a

review contains descriptions, opinions or explanations of at least one service feature or experience with the company, I consider it a high diagnosticity review. However, mere mentions of such features or experiences, without any description, opinion, justification or explanation, do not suffice for a review to be considered of high diagnosticity. I provide examples and a detailed explanation of the guidelines for high and low diagnosticity reviews in the codebook in Appendix C.

Following this approach I assign either high diagnosticity (1) or low diagnosticity (0) to each review of a random sample of 2500 reviews from the dataset. There is no specific threshold for the minimal or optimal number of texts to be annotated for supervised machine learning and natural language processing experiments. In general, the more annotated data there is, the better classifiers tend to perform. A sample size of 2500 reviews is a reasonable choice for this study, in order to build an adequate classifier and in terms of annotation workload. Six reviews were removed from the random sample because they were not written in English[9], making it impossible for me to determine their diagnosticity, thus there are 2494 annotated reviews to develop the classifier. Of those reviews about 58% (N=1445) are assigned with high diagnosticity and about 42% (N=1049) are assigned with low diagnosticity.

3.4.2 Machine Learning Classification

Machine learning approaches have been used in a broad range of research areas in the past decades. Mitchell defines machine learning as a computer program's ability to learn from experience E with respect to some class of task T and performance measure P. If its performance at task T, as measured by P, improves with experience E, then the program is said to have learned (T. M. Mitchell, 1997). The specific class of machine learning used in this study is referred to as supervised text categorization or classification. According to Sebastiani (2002), classifiers built with these approaches can achieve high levels of effectiveness, making them economically viable. This makes the approach suitable in this context, especially for large or moderately large volumes of customer reviews. I now describe how I developed my classifier by comparing several machine learning algorithms, choosing the best performer and using it to classify all reviews in my dataset (Cielen et al., 2016)[10].

I use a supervised machine learning approach with a binary classification task. This entails using the 2494 manually annotated reviews as training and testing data. I experimented

[9] Even though English reviews have been scraped, there may be some non-English reviews if reviewers have indicated their language incorrectly.
[10] My python codes can be accessed via https://github.com/ProfessorDonLuigi/ScrapeAndClassify

with several proportions between testing and training data and found the commonly used split of 20% testing (N=499) and 80% training data (N=1995) to produce the best classification performance. Prior to training the model some steps to clean the text were taken, specifically, removing symbols and unwanted characters, lowercasing, removing stop words and stemming (Sebastiani, 2002).

Feature Extraction

In order to make the unstructured text data usable for machine learning algorithms, I perform feature extraction. I use the 'term frequency - inverse document frequency' (TF-IDF) method to determine which words are most correlated with each category. Essentially, this method considers how often a word occurs in a document, compared to how often it occurs in the entire body of documents (Salton & Christopher, 1988), or applied to the present study, how often a word occurs in one review, compared to how often it occurs in the entire annotated dataset. This results in a score for each word, indicating its importance in reference to its affiliation with each category (diagnostic=1 or undiagnostic=0). In this step I also tweaked some parameters, including the number of most important features to select and found 500 to produce the best results.

Moreover, I tried different N-gram ranges, meaning that instead of just considering each word individually, word pairs are considered. This improved results for one of the tested algorithms (LinearSVC), but not the results of the best performing algorithm (Random Forest).

Classification Algorithms

There are several machine learning algorithms which can be used to perform the task of text classification. Based on different mathematical decision rules, these algorithms decide which class each document belongs to (diagnostic or undiagnostic). I tested four different algorithms and compared their classification performances to decide on the best performer. The first algorithm was Linear SVC (Support Vector Classifier) the second one was Bernoulli Naïve Bayes, the third one was Logistic Regression and the fourth one was Random Forest with the number of trees set to 100. All of these algorithms are featured in the Scikit-learn library for Python.

Performance Metrics

In order to evaluate the effectiveness of the classifier I use the common performance metrics accuracy, precision, recall and F1-score (Sebastiani, 2002). Where accuracy measures the proportion of correct predictions (i.e. correctly classified as low=0 or high=1 diagnosticity), precision measures the proportion of correctly predicted positives (i.e. proportion of reviews predicted as high diagnosticity that actually have high diagnosticity), recall indicates the proportion of actual positives which are predicted as positives (i.e. proportion of actual high diagnosticity reviews that are predicted as high diagnosticity) and F1-score represents a balance between precision and recall. Moreover, to reduce bias and increase reliability, I use 10-fold cross validation (Kim, 2009), which splits the data into 10 smaller subsets or folds. Then a model is trained on nine folds and tested on the remaining fold. This is repeated 10 times so that every fold is used as a testing set once. The performance metrics described above are then averaged across all ten iterations. I display the results for each algorithm in Table 1.

Table 1 Means of Performance Metrics from 10-fold Cross Validation

ML Algorithm	Accuracy	Precision	Recall	F1-Score
Linear Support Vector Classifier	78.17%	79.32%	85.87%	82.42%
Bernoulli Naïve Bayes	78.77%	94.45%	68.37%	79.21%
Logistic Regression	77.47%	75.44%	92.42%	83.03%
Random Forest	84.37%	89.36%	84.70%	86.59%

Choosing the Best Performer: Random Forest

Based on the performance metrics of the four tested algorithms, I consider the Random Forest algorithm to be the best performer. This algorithm performs well in all four performance metrics, scoring highest in accuracy (84.34%), second highest in precision (89.36%), second highest in recall (84.7%) and highest in F1-score (86.59%). It should be noted that deciding which particular metric is most important depends on the problem one is

trying to solve and the dataset at hand. For instance, in situations where the occurrence of one class far outweighs the occurrence of the other class, accuracy might not be the best metric to consider, since algorithms which always simply assign the majority class would score high. Clearly, such an algorithm would be of little practical use. The present dataset, however, appears to be quite balanced, since the randomly sampled 2494 reviews turned out to be 58% high diagnosticity reviews and 42% low diagnosticity reviews. Thus, it is reasonable to assume that the overall dataset does not have a strong majority class, making accuracy a meaningful metric. Furthermore, as evident in Table 1, Bernoulli Naïve Bayes and Logistic Regression achieve impeccable scores in either precision or recall respectively (>90%). If my goal was to reliably identify high diagnosticity reviews or to find as many high diagnosticity reviews as possible, I might be inclined to pick one of these two algorithms. But since my goal is to correctly classify all reviews whether diagnostic or not, Random Forest is the best choice in this context. Hence, I apply the model which was trained with the Random Forest algorithm to the entire dataset of 170,340reviews to predict the diagnosticity of each review[11]. Overall, 53.6% of the reviews are classified as diagnostic (1) according to the trained model.

3.5 Empirical Strategy

Now that all the metrics required for the analysis are computed, I explain the empirical strategy employed to estimate the effects of MRs on the quality of online reviews. I conduct a series of separate OLS regressions with each of the four review quality indicators as the outcome variable. The predictor variables are the MR characteristics as well as a set of control variables. For each quality indicator there is a baseline regression model, including the quality indicator as the outcome variable and the control variables as predictors. In the second model the MR characteristics frequency, length and speed are added as predictor variables. Each regression equation can be shown as:

$$Y = \beta_0 + \Sigma_{j=1..p}\, \beta_j X_j + \varepsilon$$

Where Y is one of the four review quality measures, β_0 is the constant of the model, X_j denotes each predictor variable, which are the MR variables plus six control variables, which are explained in section 3.5.3, and ε is the random error

In order to analyze the effects of MRs on online reviews, I split the dataset into smaller chunks of 20 reviews (N=8517) and computed averages of all variables per chunk. By splitting the dataset into chunks of 20 reviews, I intend to capture individual review pages of

[11] The model was applied to a version of the review texts which were processed in a similar way as the training and testing data, i.e. lowercased, removed unwanted characters, removed stop words and stemmed.

company profiles. As explained previously, each company profile web page displays 20 reviews at a time. Since the scraped reviews are grouped by company and chronologically ordered, this means that each chunk contains 20 consecutive reviews which were on display on the same webpage at the moment of data collection. In other words, each chunk represents one webpage of a company profile. Thus, this setup is designed to measure effects of MRs on reviews within one company profile webpage. It is difficult to gauge exactly how many reviews customers actually read before having sufficient information about a company or product. This likely depends on several characteristics such as type of product/service, price, prior experience with the company or economic standing of the consumer. According to a yearly survey conducted by BrightLocal, consumers tend to read between seven and thirteen reviews, depending on age group, before deciding whether to trust a business (Murphy, 2019). However, these are not the only reviews which affect consumers, since consumers require up to 40 reviews to be displayed, in order to deem the presented information credible (Murphy, 2018). Thus, grouping reviews as chunks of twenty is a realistic approximation of how many reviews consumers are likely exposed to on average. This is also sensible in the case of Trustpilot reviews, considering that 20 reviews are displayed on one company profile website at any time.

Concerns associated with the fact that the type of company may affect review quality metrics are addressed by the selection of companies within one business category, which, according to Trustpilot, lists comparable businesses. Furthermore I include some company specific control variables, which are discussed below.

3.5.1 Independent Variables: MR Frequency, Length & Speed

The three independent variables in my framework are MR frequency, MR speed and MR length. I will now explain how each of these concepts is measured.

Following previous studies (Chevalier et al., 2018; Lui et al., 2018; Schuckert et al., 2019), I measure MR frequency as the ratio of reviews which received a response to total reviews. MR Speed is measured as the difference in days between posting of a review and posting of the response (Sheng, 2019). Finally, I measure MR length as the word count of the response text (Sheng, 2019).

3.5.2 Dependent Variables: Review Quality Measures

The dependent variable is review quality. Review quality is determined by four individual measures representing length, useful votes, diagnosticity and readability. First, I

consider the length of a review as a quality indicator, with higher length indicating higher quality (e.g. Burtch et al., 2018; Khern-am-nuai et al., 2018). Again, length is simply measured by counting the number of words. In order to reduce skewness, review length was log-transformed (Cielen et al., 2016). Second, I consider the number of useful votes on a review to be an indicator of quality with more useful votes on a review indicating higher quality (e.g. Khern-am-nuai et al., 2018; Li et al., 2017). Like the previous measure, the number of useful votes was log-transformed, in order to reduce skewness with the addition of a small constant, to deal with cases where the value is zero (Cielen et al., 2016). Third, to measure diagnosticity, I make use of the binary machine learning classifier (section 3.4), which indicates if product or service specific features are present in a given review or not. If such features are present, it is an indicator for higher quality. The fourth measure is the readability index, comprised of several common readability formulas as discussed in section 3.3. As in previous studies on online reviews, higher readability scores indicate higher review quality (e.g. Khern-am-nuai et al., 2018; Korfiatis et al., 2008; Li, Zhang, Janakiraman, & Meng, 2016).

3.5.3 Control Variables

I make use of six control variables, which may affect the quality of online reviews. First, review rating has been shown in previous studies to have an effect on consumer reviewing behavior, for instance, Chevalier et al. (2018) show that reviews with lower rating tend to be longer. In the same study, the authors also demonstrate that rating can affect perceived helpfulness of reviews. In the context of the present study, this is especially pertinent to the number of useful votes on reviews and the diagnosticity of reviews.

Second, to control for the linguistic level of the responses, I include the readability score of MRs. This is measured in the same way as for review readability as described in section 3.3. Observing the linguistic level of responses to reviews of other users might affect users in writing their own review, since they might expect a certain linguistic level in the response to their review.

Third, I control for the number of reviews per company. The volume of reviews is prominently displayed on each Trustpilot company profile and may affect consumers' reviewing behavior (Li et al., 2017).

Fourth, I control for the status of the company profile on Trustpilot, indicating whether the company profile has the status 'Asking for Reviews' or 'Claimed' with a dummy variable. Again, this attribute is visibly on display at the top of each company profile and may affect

consumers reviewing behavior. Moreover, it is conceivable that businesses which actively collect reviews have more engaged customers posting reviews. Thus, it is reasonable to control for the status of the company profiles.

Fifth, I consider whether a review has been verified as a genuine purchase or experience. Reviews are marked as verified if they are written upon a company's invitation, or if documentation of a buying or service experience is provided by the reviewer[12]. Review quality may be affected by either of those two circumstances.

Finally, I control for the total number of reviews a user has posted on Trustpilot. Users who have posted multiple Trustpilot reviews can be considered more active, engaged and experienced users of the platform. This may have an effect on the quality of their contributions. In Table 2, I list all dependent, independent and control variables with a brief description.

Table 2 Variable Descriptions

Variable	Description
Review Quality Variables (dependent Variables)	
Review Length	Number of words in the review text
Useful Votes	Number of useful votes a review received
Readability Review	Readability score of reviews based on readability formula consensus
Diagnosticity	Denotes whether a review is diagnostic or not, 1 = diagnostic 0 = not diagnostic
MR Variables (independent variables)	
MR Frequency	Share of reviews which received a response
MR Length	Number of words in the response text
MR Speed	Number of days between posting of review and posting of response, 0 = within same day, 1 = within one day, 2 = within two days, etc.
Control Variables	
Rating	Review rating between 1 and 5 stars
Status	Status of company profile, 1 = Asking for Reviews 0 = Claimed
Review Volume	Total number of English reviews on company profile
Verified	Denotes whether a review is verified as an actual consumer experience, 1 = verified 0= not verified
User Reviews	Number of reviews a user has written
MR Readability	Readability score of response based on readability formula consensus

Table 3 displays the summary statistics for the dependent, independent and control variables including the means, standard deviations, minima and maxima. One of the statistics revealed in the tables is that about 15.56% of the reviews in the dataset received a response

[12] https://support.trustpilot.com/hc/en-us/articles/201819697-Why-are-some-reviews-marked-Verified-

from the company. Apart from the frequency with which reviews receive responses, the table shows that the average MR length is at about 39 words. The longest response in the dataset is 737 words long. When it comes to response speed, it can be seen that the fastest responses were posted within the same day, and the slowest were posted within nine days. On average responses were posted between one and two days after the review (1.609).

Furthermore, Table 3 provides some insight into the overall quality characteristics of the reviews in the dataset. It can be seen that, on average, reviews are about 37 words long. However, the range is quite large with the longest review being 1840 words in length. The mean number of useful votes is at 0.039, which indicates that most reviews in the dataset do not receive helpful votes. The maximum number of useful votes on one review is 11. Readability scores of the reviews average close to a nine, suggesting that a person with nine years of English education should understand the text upon reading it once (DuBay, 2004). The readability scores reach some extreme values at the top end (max 788) and cannot sensibly be interpreted with the same logic in those cases. These values can occur, for instance, if reviews are written without the use of spaces in the text, due to the way in which readability scores are calculated. Finally, we can see that the ratio between diagnostic and undiagnostic reviews is quite even, with 53.6% of reviews being diagnostic.

Table 3 Summary Statistics

	Count (Share)	Mean	Std. Dev.	Min	Max
Reviews	170,340 (100%)				
Responses (Overall Response Share)	26,508 (15.56%)				
Review Length	170,340	37.478	47.267	1	1840
Useful Votes	170,340	0.039	0.300	0	11
Readability Review	170,340	8.859	6.081	0	788
Diagnosticity	170,340	0.536	0.499	0	1
Rating	170,340	3.881	1.556	1	5
Status	170,340	0.83	0.37	0	1
Review Volume	170,340	3097.09	8173.15	500	60420
Review is Verified	170,340	0.716	0.451	0	1
User Reviews	170,340	1.378	0.931	1	9
MR Readability	26,508	9.44	7.03	0	105
Response Length	26,508	39.193	43.686	1	737
Response Speed in Days	26,508	1.609	1.923	0	9

Table 4 shows the bivariate correlations between all variables. Due to some high correlations among some of the variables, I calculated the variance inflation factors (VIF) to check for multicollinearity. As apparent in Table 5, all variables have VIF scores below 5, thus, no multicollinearity issue is detected (Schroeder, Lander, & Levine-Silverman, 1990).

Table 4 Bivariate Correlations between Variables grouped Chunks of 20 reviews (N=8517)

* = p<0.1, ** = p<0.05, *** = p<0.01

	1	2	3	4	5	6	7	8	9	10	11	12	13
1 Review Length (log)	1												
2 Useful Votes (log)	0.23***	1											
3 ReadReview	0.29***	0.05***	1										
4 Diagnosticity	0.80***	0.20***	0.28***	1									
5 MRFreq	-0.16***	0.09	0.02	-0.18***	1								
6 MRLength	0.15***	-0.09***	0.16***	0.14***	-0.40***	1							
7 MRSpeed	-0.11***	0.01	-0.07***	-0.13***	0.05***	-0.03**	1						
8 MRRead	0.11***	-0.11***	0.07***	0.15***	-0.34***	0.35***	-0.1***	1					
9 Rating	-0.43***	-0.03***	-0.08***	-0.29***	0.29	-0.01	0.02	0.05***	1				
10 Status	-0.11***	-0.32***	-0.02**	-0.1***	-0.4***	0.29***	-0.15***	0.29***	-0.23	1			
11 Verified	-0.11***	-0.36***	-0.04*	-0.17***	-0.4***	0.28***	-0.03	0.22***	-0.38***	0.63***	1		
12 UserReviews	-0.18***	0.09***	-0.00	-0.17***	0.26***	0.01	0.06***	-0.03*	0.33	-0.13*	-0.14**	1	
13Review Volume	0.29***	-0.21***	0.02	0.16***	-0.38**	0.16***	-0.08***	0.20***	-0.83***	0.34	0.49	-0.43	1

Table 5 Variance Inflation Factor Scores

Variable	VIF-Score
Constant	383.05
Review Length (log)	3.45
Useful Votes (log)	1.58
Readability	1.16
Diagnosticity	3.19
MR Frequency	1.79
MR Length	1.37
MR Speed	1.06
Readability Response	1.30
Rating	1.58
Status	2.39
Verified	2.95
User Reviews	1.05
Review Volume	1.21

4 Results

In Tables 6-9, the regression results for the four outcome measures review length, useful votes, readability and diagnosticity are displayed with a description and the implications for the corresponding hypotheses under each table. Each table contains the base model and the model which includes the MR variables frequency, length and speed. I also include the adjusted R squared for each model which indicates the variance in the dependent variable, which can be explained by the model. Comparing the adjusted R squares of the base models to the models which include the MR measures, provides some insight into how much added variance is explained by the MR variables.

Table 6 Regression Results for Review Length (log)

Variable	**Base Model**			**MR Model**		
	b	s.e.	p	b	s.e.	p
MRFreq				-0.226	0.020	0.000
MRLength				0.001	0.000	0.000
MRSpeed				-0.020	0.004	0.000
MRRead	0.010	0.001	0.000	0.004	0.001	0.000
Rating	-0.421	0.013	0.000	-0.408	0.012	0.000
Status	0.108	0.021	0.000	0.057	0.021	0.006
Verified	-0.419	0.020	0.000	-0.501	0.021	0.000
UserReviews	-0.154	0.018	0.000	-0.135	0.018	0.000
Review Volume	-0.000	-0.000	0.000	-0.000	-0.000	0.000
Constant	5.406	0.062	0.000	5.531	0.061	0.000
Adj. R squared	0.378			0.425		

First, I consider hypotheses 1a, 2a and 3a, which all pertain to effects on review length. The base model and MR model are shown above, in Table 6. Contrary to hypothesis 1a, the analysis reveals a negative and significant effect of MR frequency on review length (-0.226, $p<0.001$). For MR length, there is a positive and significant effect on review length (0.001, $p<0.001$), confirming hypothesis 2a. Furthermore, the analysis reveals that faster MR speed positively affects review length (-0.020, $p<0.001$), which is in line with hypothesis 3a. Note that negative coefficients for review speed indicate that shorter time spans between review and response lead to increases in the outcome variable. Comparing the adjusted R squares of the base model and the MR model, there is an increase from 37.2% to 42.5%.

Table 7 Regression Results for Useful Votes (log)

Variable	Base Model			MR Model		
	b	s.e.	p	b	s.e.	p
MRFreq				-1.613	0.097	0.000
MRLength				-0.001	0.001	0.035
MRSpeed				-0.016	0.020	0.412
MRRead	0.002	0.005	0.646	-0.016	0.005	0.001
Rating	-1.626	0.060	0.000	-1.609	0.057	0.000
Status	-0.827	0.099	0.000	-0.964	0.098	0.000
Verified	-1.026	0.097	0.000	-1.508	0.099	0.000
UserReviews	-0.189	0.087	0.030	-0.072	0.084	0.390
Review Volume	-0.000	-0.000	0.016	-0.000	-0.000	0.040
Constant	2.986	0.294	0.000	4.029	0.292	0.000
Adj. R squared	0.301			0.351		

Second, I present the results for useful votes in Table 7, which relate to hypotheses 1b, 2b and 3b. MR frequency has a negative and significant effect on the number of useful votes (-1.613, $p<0.001$) leading me to refute hypothesis 1b. The effect of MR length on the number of useful votes is negative and significant (-0.001, $p<0.035$), leading me to refute hypothesis 2b as well. For MR speed no significant effect was found (-0.016, $p=0.412$), thus hypothesis 3b is not supported. Comparing the adjusted R squares of the base model and the MR model, there is an increase from 30.1% to 35.1%.

Table 8 Regression Results for Readability

Variable	Base Model			MR Model		
	b	s.e.	p	b	s.e.	p
MRFreq				0.200	0.080	0.013
MRLength				0.005	0.001	0.000
MRSpeed				-0.054	0.016	0.001
MRRead	0.016	0.004	0.000	0.008	0.004	0.048
Rating	-0.594	0.048	0.000	-0.564	0.047	0.000
Status	0.259	0.080	0.001	0.172	0.080	0.033
Verified	-0.291	0.078	0.146	-0.278	0.081	0.001
UserReviews	0.003	0.070	0.965	-0.003	0.069	0.964
Review Volume	-0.000	-0.000	0.146	-0.000	-0.000	0.983
Constant	11.276	0.236	0.000	11.0824	0.241	0.000
Adj. R squared	0.051			0.074		

Third, the regression results for review readability are illustrated in Table 8. These apply to hypotheses 1c, 2c and 3c. Readability is shown to be positively affected by MR frequency (0.200, $p=0.013$), which is in line with hypothesis 1c. Higher MR length has a positive effect on readability (0.005, $p<0.001$), confirming hypothesis 2c. Similarly, higher MR speed has a positive effect on readability score (-0.054, $p<0.001$), confirming hypothesis 3c. Comparing the adjusted R squares of the base model and the MR model, there is an

increase from 5.1% to 7.4%. It is evident that, more so than for the other outcome variables, there are other important factors causing the variation in readability which are not captured in the models.

Table 9 Regression Results for Diagnosticity

Variable	Base Model			MR Model		
	b	s.e.	p	b	s.e.	p
MRFreq				-0.128	0.009	0.000
MRLength				0.0004	-0.000	0.000
MRSpeed				-0.011	0.002	0.000
MRRead	0.0053	0.000	0.000	0.003	0.000	0.000
Rating	-0.1593	0.006	0.000	-0.154	0.005	0.000
Status	0.0643	0.009	0.000	0.040	0.009	0.000
Verified	-0.1755	0.009	0.000	-0.216	0.009	0.000
UserReviews	-0.0989	0.008	0.000	-0.088	0.008	0.000
Review Volume	-0.000	-0.000	0.000	-0.000	-0.000	0.000
Constant	1.289	0.028	0.000	1.371	0.027	0.000
Adj. R squared	0.332			0.387		

Finally, results for diagnosticity are shown in Table 9, these pertain to hypotheses 1d, 2d and 3d. Effects of MR frequency on review diagnosticity are shown to be negative (-0.128, $p<0.000$), thus hypothesis 1d is refuted. For MR length, on the other hand, a positive effect on review diagnosticity is observed (0.0004, $p<0.001$), confirming hypothesis 2d. MR speed is also shown to positively affect review diagnosticity, confirming hypothesis 3d (-0.011, $p<0.001$). Comparing the adjusted R squares of the base model and the MR model, there is an increase from 33.2% to 38.7%.

By inspecting the base models, some insights are provided about which of the control variables have stronger and weaker impacts on changes in the outcome variables. For instance, *Rating* and *Verified* appear to have quite strong impacts on the outcome variables per unit of change in all models. When inspecting the coefficient of *Review Volume*, however, it appears that the impact is very small across all models. Note that the coefficient indicates changes in the outcome variable caused by a one unit change of the explanatory variable. In the case of *Review Volume*, this represents a change of one review posted on a company profile, which is a very minor change considering that some profiles have over 10,000 reviews. Thus, it should not automatically be concluded that the variable is not a good predictor in the models, it is however worth investigating. Therefore, I re-ran the regressions without the variable *Review Volume* and found that the R squared values did in fact decrease, especially for the review length and diagnosticity models. The coefficients of the indicator

variables all remained equal in direction, with slight changes in strength[13]. Consequently, I decided to keep *Review Volume* in the analysis.

5 Discussion & Conclusion

5.1 Main Findings

On many online review platforms, companies can respond to the feedback provided to them by customers. Some companies make use of this feature far more frequently and extensively than others. In this study, I set out to answer the question: *'What is the effect of managerial responses on the quality of online customer reviews?'* In particular, I examined the effects of the frequency with which reviews are answered, the response length and the response speed. By combining web scraping, machine learning, natural language processing and regression analysis, a large dataset of online reviews and responses was retrieved and analyzed. Four individual review quality measures were considered, the length of online reviews, the number of useful votes, the readability and the diagnosticity. The analysis has revealed that higher MR frequency is not effective in improving the quality of online reviews, while higher MR speed and MR length do improve the quality of online reviews. In the following, I discuss the main findings, their practical implications, theoretical contributions as well as limitations and avenues for future research.

5.2 Key Indicators of Review Quality: Review Length & Diagnosticity

The preceding analysis has shown that the four quality measures are affected in different ways by the examined MR aspects. Higher response frequency, for instance, only has a positive effect on the readability of reviews, while review length, useful votes and diagnosticity are affected unfavorably by higher MR frequency. For MR length, the analysis suggests favorable effects on review length, readability and diagnosticity. Only the number of useful votes was affected negatively. Higher MR speeds produced the most unambiguously positive effects on review quality, with positive effects on review length, readability and diagnosticity and a non-significant effect on the number of useful votes.

With the use of several review quality indicators, it is beneficial to look at each indicator against the background of the conducted analysis and in the context of the present dataset. This is useful in order to be able to grasp which of the indicators could be considered more or less important, specifically if one intends to generalize to other contexts and to make

[13] The regression tables of the analysis without review volume are in Appendix D.

recommendations to companies and other actors. So, which of the quality indicators should be considered most important here?

The four quality indicators were selected, in part, based on their (common) usage in previous studies on online review quality. Throughout the analysis in this study, it has become evident why it is important to incorporate several measures of review quality, since by themselves they have certain shortcomings. For instance, measuring review quality only by usefulness votes can be problematic, if most reviews do not receive any votes. This is certainly the case in the context of Trustpilot reviews. Out of more than 170,000 reviews in the dataset only about 4,200 have received any useful votes. Thus, it would not be adequate to judge review quality solely based on this metric. I would still consider useful votes insightful, however, not quite as crucial as the other metrics.

In the case of readability scores, certain weaknesses can arise if reviews are very short but use complex words. In some cases, this can lead to highly inflated scores even if the actual utility of a review is quite low. This problem was addressed by using a consensus of several readability formulas, but some rather inflated readability scores still occurred. Additionally, as evident in the determination coefficient of the readability regression model (adj. r2 = 0.074), there are factors which are not incorporated in the present analysis, strongly affecting readability scores of online reviews. Again, I still consider readability score a useful metric. However, results should be interpreted carefully and with consideration for the texts which are being analyzed. In this dataset, reviews are about 37 words long on average, many are even shorter than 20 words. As reviews get shorter, readability can become a more volatile and less reliable quality indicator (DuBay, 2004).

In my assessment, the most useful review quality measures for the present dataset are the review length and the diagnosticity classifier developed in this study. Review length may be a simple measure but often it is a good indicator of the utility and information richness of an online review. This is particularly true in a context where reviews tend to be rather short. The usage of the diagnosticity concept, adds nuance by focusing on the inclusion of product or service specific information. These types of information are instrumental for review readers in order to make informed purchase decisions. So to come back to which review quality metrics are most important here, I consider the length and diagnosticity to be the most powerful and adequate measures followed by readability and useful votes.

5.3 Practical Implications

Against the background of the discussion and explanation about which quality indicators I deem most important, I can now make well founded recommendations in reference to MR frequency, length, and speed. Firstly, it is quite clear, based on the analysis presented here that organizations should try to respond to reviews as quickly as possible. Review length, diagnosticity and readability can be increased by responding in a speedy manner. These findings are consistent with some of the literature reviewed in this study. Faster responses signify that a company takes its customers' concerns seriously (Sparks et al., 2016) and also create an appearance of active embrace of customer comments (Sheng, 2019). Perceptions of a responsive management and the prospect of receiving responses quickly appear to translate into higher quality reviews. In practice it will be difficult to always respond to reviews within one or two days of posting, especially as review volumes of organizations increase. But if review volumes are manageable, it is worth investing resources into responding within just a few days.

The length of MRs is shown to be a useful tool in inducing higher quality reviews as well. In particular, review length, diagnosticity and readability can be improved by giving longer MRs. Based on uncertainty reduction theory (Daft & Lengel, 1984), longer MRs are better suited to convey information, reduce ambiguity and in turn facilitate the purchasing process. This is again a powerful means of displaying care for one's customers and dedication to address and solve their concerns. Thus, I recommend that when customer support professionals, managers or other company representatives decide to answer a review, they should do so in a detailed and informative manner, such that the addressed user, and preferably also the other users who may read the exchange, face reduced uncertainty. It should be noted though that, in practice, there surely is a limit to the benefit of longer MRs. Clearly, it is not advisable to keep adding verbose sentences to the response just for the sake of posting longer responses. MRs should still be to the point and concise, but if relevant additional points can be provided, this may offer a range of benefits to the customers and the company, including improvements in the quality of online reviews.

Finally, I consider the frequency of MRs. Contrary to the hypotheses developed in regards to MR frequency, higher MR frequencies did not improve the quality of online reviews. The only exception to this was an improvement in review readability, however, all the other quality metrics were impacted negatively by higher MR frequency. This seems to go against the expectation that a display of customer care and engaged communication, to which I attributed the effectiveness of the two other MR aspects, improves review quality. Following

the same logic one should expect a higher share of responses to translate into higher review quality. However, there appear to be other unintended effects which cause higher MR frequency to impair the quality of online reviews. One way in which a high MR frequency might manifest in lower review quality on a reviewing platform like Trustpilot could be that responding to too many reviews incentivizes undesirable reviewing behavior. I use the word undesirable here specifically from a perspective of online review quality. If we assume that it is desirable for review writers to receive a response and they see that a company only responds to some reviews, review writers may be inclined to present a well-stated and well-reasoned case, so that they are more likely to get a response. On the other hand, if a company responds to a majority or even all reviews, regardless of quality, reviewers might not feel the need to write a particularly high quality review. By this logic, it may be beneficial to be somewhat more selective in choosing which reviews to respond to. Responding predominantly to detailed, well explained and well written reviews may signify that reviewers are more likely to be 'awarded' a response, if the review warrants one.

In combination with the findings regarding MR speed and length, there are some useful and actionable takeaways from the realization that higher MR frequency does not benefit the quality of online reviews. The findings imply that, rather than striving to answer as many customer reviews as possible, it is more beneficial to respond to a smaller amount of reviews and to do so as timely and thoroughly as possible. In practical contexts, companies have limited resources in terms of time and personnel, which can be expended towards the task of webcare. This study has shown that using these resources selectively is more beneficial than attempting to maximize coverage. Company representatives, who work at the task of responding to customer reviews, can implement these findings in several ways. First, they can try to respond to customer reviews on a regular basis and in short intervals. Responding to a small number of reviews every day, rather than a large number of reviews once a week ensures that response speed remains consistently high. Second, they can select reviews which are detailed, informative and well written and respond to them in a manner which is detailed and insightful for the reviewer and potentially even other users. Thus, high review quality in subsequent reviews can be induced.

Of course it will not always be trivial for employees to reliably spot which reviews are the highest quality reviews. I have shown, for instance, that the number of useful votes may not always effectively aid in spotting high quality reviews, since so few reviews actually receive useful votes. Similarly, assessing reviews based on readability score upon reading them seems neither convenient nor very reliable, especially for shorter texts. Review length

and diagnosticity, on the other hand, represent review quality more visibly and more practically. Using these two concepts as guidelines might be a good basis for selecting which reviews to respond to. Alternatively, as I have shown in this study, the process of detecting high or low diagnosticity reviews could even be automated. Following this approach, a pre-selection of potentially response-worthy reviews could be made automatically. Subsequently, company representatives could choose which reviews they actually want to respond to.

5.4 Implications for Review Platforms

This study also bears some implications from the perspective of review platforms such as Trustpilot, although the results can be generalized to other platforms as well, since they share functionalities and features, which are present on Trustpilot. It has become evident that the feature of up-voting reviews as 'useful' is utilized quite rarely. Furthermore, it can be seen that many reviews on the platform are quite short. Currently, the threshold for a review to be posted on Trustpilot is eleven characters, which may be convenient to many reviewers if they do not wish to exert a lot of effort but still want to give some form of feedback. It is difficult to imagine, however, that a customer review can be truly insightful at such a low limit. So while a limit like this may boost the number of reviews posted on the website, it may deteriorate the quality of online reviews.

Another noteworthy aspect which was revealed in the analysis is related to the *Verified* control variable, which denotes whether a review was written upon a company's invitation to review them. This variable had a rather pronounced negative effect on all four review quality indicators, which would suggest that a review invitation substantially influences the quality of reviews in a negative direction. It is not immediately obvious why such an invitation should have this effect but this is certainly an interesting phenomenon to be investigated. One possible explanation is that many of the invited reviewers may have received some type of compensation for contributing their review, in the form currency or company credit. This practice has been shown to undermine perceived trustworthiness of online reviews by boosting review quantity and in some cases rating score, while simultaneously causing review quality to deteriorate (Burtch et al., 2018; Khern-am-nuai et al., 2018). However, it is not trivial to test the impact of compensation for review writing empirically. This is because many countries prohibit offering reimbursements for reviews[14] [15], which in turn means that companies do not necessarily disclose whether they reimburse review writers publicly.

[14]https://www.ftc.gov/news-events/press-releases/2009/10/ftc-publishes-final-guides-governing-endorsements-testimonials

[15] https://www.it-recht-kanzlei.de/gegenleistung-kundenbewertung-online-shop.html

Perhaps the negative impact of financial incentives on review quality accounts for some of the effects of *Verified* reviews, observed in this study.

The aspects of perceived review integrity and impartiality are central to review platforms and to online reviews in general (Korfiatis, García-Bariocanal, & Sánchez-Alonso, 2012; Liang et al., 2017). If the displayed reviews are of high volume and rating but of very low quality, customers' trust into the platform may eventually erode severely. This concern is becoming more urgent, since systematic posting and acquisition of fraudulent reviews have recently occurred more often (Lappas, Sabnis, & Valkanas, 2016; Luca & Zervas, 2016). Simultaneously, consumers are becoming more vigilant in this matter. According to a survey about online reviews on local businesses in the UK by Murphy (2018), 74 percent of consumers indicated that they had read fake reviews in the preceding year. It is difficult to combat online review fraud and abusive tactics, since, even if legislation may exist[16], it can be difficult to enforce it effectively. Against this background, higher quality reviews induced by the correct use of MRs, may provide some remedy, since they invoke more confidence in the authenticity of reviews.

To address some of the problems mentioned above, review platforms might consider some form of gamification, for instance by introducing a point system where users are awarded virtual points or ranks for being active and making high quality contributions (Xu et al., 2020). Alternatively, some of the best reviews could be displayed prominently as influential or important reviews on a company profile, to reward users who make excellent contributions and to provide an additional quality indicator for review readers (Xu et al., 2020).

Finally, a beneficial addition to the functionality of Trustpilot, and reviewing platforms in general, could be a feature which lets users filter for reviews which have received a response. Currently, such a filtering option exists for different star ratings awarded by the reviewers. When considering that responding to a large portion of reviews is not recommendable for companies, this would provide an opportunity to give additional weight and exposure to high quality reviews, or reviews which companies deem response worthy. For users, this may be a convenient way to grasp how a company communicates with its customers.

[16] §5 Abs. 2 Satz 1 Nr. 1 UWG Article of German Competition law which legislates unfair competition

5.5 Theoretical Contributions

The findings of the present study bear several theoretical implications. First, I have demonstrated that the domain of MRs, which has thus far mainly been researched in hospitality and tourism contexts, can be extended to other e-commerce contexts. It has been suggested in previous studies, that generalizations about online reviews and MRs are not limited to businesses like hotels, restaurants and travel providers (Chen et al., 2019), this assertion has been tested and verified in the present study. Companies which focus on online services, such as online communications, web hosting, IT consulting and VPN services were the center of attention in this study. The findings provided by this study contribute to the present MR literature by adding empirical evidence pertaining to the effects of MR frequency, speed and length, which had previously been deemed advisable by some authors (Chevalier et al., 2018; Xu et al., 2020) and ineffective by others (C. Li et al., 2017).

Moreover, this study demonstrates the utility in combining several methods including web scraping (R. Mitchell, 2018), machine learning (T. M. Mitchell, 1997), natural language processing (Sebastiani, 2002) and regression analysis. By developing an automatic review diagnosticity classifier, I incorporated a text based review quality measure, which is better suited to capture the content of online reviews than most of the previously used review quality measures. Furthermore, the combination of these methods enabled the retrieval and analysis of a large dataset of online reviews. Especially in previous studies about MRs, machine learning and natural language processing methods have been employed quite rarely.

Another crucial contribution provided by this study is the investigation of the review platform Trustpilot, which, to the best of my knowledge has not yet been the subject of empiric analysis with large amounts of review data[17]. Trustpilot is an interesting research context since it allows users and researchers to observe an immense range of industries and organizations. On the one hand, business-to-customer interactions can be studied, by examining how firms communicate with their customers, on the other hand, the platform provides a glimpse into how closely companies cooperate with Trustpilot, which can be considered a business-to-business relationship. Some of these business-to-business interactions include whether companies make use of Trustpilot's review invitation feature, whether a company profile is actually affiliated with the company or whether a company reports abusive or fraudulent reviews for Trustpilot to censor and remove. From an information systems research perspective, it is interesting to study online spaces in which

[17] The only exception I found is a study about deceptive reviews and spam reviews (Sandulescu & Ester, 2015)

companies are not in full control of the content displayed on their profile. Businesses do not have the option to choose which customer reviews and interactions are on display and which ones are censored. This competence, in the case of Trustpilot, lies with the moderators of the platform. Moreover, filtering and selecting which reviews to read typically lies with the user. (Khern-am-nuai et al., 2018). For instance, if users wish to filter for poor reviews, they are free to do so. I have shed light on how organizations can interact with customers on a third party website, in order to create a favorable image, even though they do not retain full control of this online space. The feature of responding to customer feedback is available on many third party websites, thus, the findings about MR frequency, speed and length can be generalized to other online spaces of discourse between organizations and customers.

5.6 Limitations and Future Research

There are some limitations present in this research, which should be addressed in future studies. One of them pertains to the fact that certain aspects of MRs have not been considered in the empirical analysis. Specifically, aspects related to the content of MRs have not been considered in great detail. However, the topics and tone of voice may be crucial to the efficacy of MRs in inducing high quality reviews. In the future, qualitative characteristics of MRs and their effects on the quality of reviews should be investigated more thoroughly. Methods such as topic modeling (e.g. Latent Dirichlet Allocation) could be used to address tasks like extracting latent themes within review texts. These approaches could facilitate the incorporation of qualitative features in reviews and MRs on large datasets. This may include individual topics in MRs (Xie et al., 2017) which might affect subsequent review quality, the tone of voice used in MRs, for instance, defensive, grateful, accommodative or apologetic tones and their effects (Weitzl, 2019) or the degree to which a response matches the content of the review (Roozen & Raedts, 2018; X. Zhang et al., 2020).

Second, I used traditional natural language processing and machine learning algorithms to develop an automatic text classifier. These are certainly not the most cutting edge methods in the field of machine learning. Artificial neural networks and deep learning might be applied to achieve better predictive performance than the methods used in this study.

Third, other empirical strategies could be pursued. For instance, structural equation modeling may be insightful in the context of review quality, since multiple indicator variables were used to explain the overarching construct of review quality. Moreover, additional control variables, which might improve the determination coefficients of the models, can be incorporated. For instance, I have briefly discussed the idea that financial incentives may play

an important role in explaining the quality of online reviews. For this, other types of data would have to be collected and combined with the present dataset or comparable datasets. Additionally, future study designs may want to address concerns related to possible changes in examined companies over time. In this study, reviews from a timespan of several years were scraped. Presumably, some companies underwent major changes in this period, which may have affected the online review quality.

Fourth, I have addressed some of the shortcomings of measures such as readability formulas, especially when applied to very short texts. Results for this indicator should be interpreted with care and consideration for the texts which are being analyzed. Even diagnosticity, as measured in this study, is subject to some potential biases. Although the automatic classifier performed well, in a dataset of over 170,000 online reviews a considerable number of reviews should still be expected to be classified incorrectly. More fundamentally, it should be noted that the classifier is based on human input provided via the annotated data. The annotated reviews were only annotated by one person. It would be desirable to have additional coders, so that inter coder reliability can be calculated. However, due to the scope of the project, it was not feasible to enlist additional coders.

Finally, it may be interesting to compare review quality across different reviewing platforms. Conceivably, the layout and functionality of a platform can affect the quality of online reviews, for instance, by enforcing different character or word count thresholds for reviews to be submitted. In future studies, Trustpilot can be incorporated in such cross platform comparisons.

6 Bibliography

Alrawadieh, Z., & Dincer, M. Z. (2019). Reputation management in cyberspace: evidence from Jordan's luxury hotel market. *Journal of Hospitality and Tourism Technology*, *10*(1), 107–120. https://doi.org/10.1108/JHTT-09-2017-0093

Burke, V., & Greenberg, D. (2010). Determining Readability: How to Select and Apply Easy-to-Use Readability Formulas to Assess the Difficulty of Adult Literacy Materials. *Adult Basic Education and Literacy Journal*, *4*(1), 34–42.

Burtch, G., Hong, Y., Bapna, R., & Griskevicius, V. (2018). Stimulating Online Reviews by Combining Financial Incentives and Social Norms. *Management Science*, *64(5)*(February 2020), 2065–2082.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, *50*(2), 511–521. https://doi.org/10.1016/j.dss.2010.11.009

Casado-Díaz, A. B., Andreu, L., Beckmann, S. C., & Miller, C. (2020). Negative online reviews and webcare strategies in social media: effects on hotel attitude and booking intentions. *Current Issues in Tourism*, *23*(4), 418–422. https://doi.org/10.1080/13683500.2018.1546675

Chen, W., Gu, B., Ye, Q., & Zhu, K. X. (2019). Measuring and Managing the Externality of Managerial Responses to Online Customer Reviews. *Information Systems Research*, *30(1)*(November), 81–96.

Chevalier, J. A., Dover, Y., & Mayzlin, D. (2018). Channels of Impact : User Reviews When Quality Is Dynamic and Managers Respond. *Marketing Science*, *37(5)*(December 2019), 688–709.

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing data science : big data, machine learning, and more, using Python tools*.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284. https://doi.org/10.1037/h0076540

Daft, R. L., & Lengel, R. H. (1984). Information Richness: A New Approach to Managerial Behavior and Organization Design. *Research In Organizational Behavior*, *6*, 191–233. https://doi.org/N00014-83-C-0025

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, *27*(2), 37–54. https://doi.org/10.2753/JEI0021-3624440403

Davidow, M. (2003). Organizational Responses to Customer Complaints: What Works and

What Doesn't. *Journal of Service Research*, *5*(3), 225–250. https://doi.org/10.1177/1094670502238917

De Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *Journal of Interactive Marketing*, *26*(2), 83–91. https://doi.org/10.1016/j.intmar.2012.01.003

Dennis, A. R., Fuller, R. M., & Valacich, J. S. (2008). Media, tasks, and communication processes: A theory of media synchronicity. *MIS Quarterly: Management Information Systems*. https://doi.org/10.2307/25148857

DuBay, W. (2004). The Principles of Readability. *Online Submission*, (January 2004).

Fan, J., & Li, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. Retrieved from http://arxiv.org/abs/math/0602133

Feng, W., & Ren, W. (2019). "This is the destiny, darling": Relational acts in Chinese management responses to online consumer reviews. *Discourse, Context and Media*, *28*, 52–59. https://doi.org/10.1016/j.dcm.2018.09.003

Flesch, R. (1979). How to write in plain English: A book for lawyers and consumers. *New York: Harper*.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Information Systems Research*, *19*(3), 291–313. https://doi.org/10.1287/isre.1080.0193

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188

Goes, P. B., Lin, M., & Yeung, C. man A. (2014). "Popularity effect" in user-generated content: Evidence from online product reviews. *Information Systems Research*, *25*(2), 222–238. https://doi.org/10.1287/isre.2013.0512

Goh, K., Heng, C., & Lin, Z. (2013). Social Media Brand Community and Consumer Behavior : Quantifying the Relative Impact of User- and Marketer- Generated Content. *Information Systems Research*, *24(1)*(February 2020), 88–107.

Gunning, R. (1952). The technique of clear writing. *New York: McGraw-Hill.*

Gyasi, W. K. (2013). Readability and Academic Communication: A Comparative Study of Undergraduate Students' and Handbook of Three Ghanaian Universities. *IOSR Journal of Computer Engineering*, *13*(6), 41–50. https://doi.org/10.9790/0661-1364150

Ho, V. (2017). Achieving service recovery through responding to negative online reviews.

Discourse and Communication, *11*(1), 31–50. https://doi.org/10.1177/1750481316683292

Khern-am-nuai, W., Kannan, K., & Ghasemkhani, H. (2018). Extrinsic versus Intrinsic Rewards for Contributing Reviews in an Online Platform. *Information Systems Research*, *29(4)*(February 2020), 871–892.

Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, *53*(11), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009

Kim, S. J., Wang, R. J. H., Maslowska, E., & Malthouse, E. C. (2016). "understanding a fury in your words": The effects of posting and viewing electronic negative word-of-mouth on purchase behaviors. *Computers in Human Behavior*, *54*, 511–521. https://doi.org/10.1016/j.chb.2015.08.015

Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. *COLING/ACL 2006 - EMNLP 2006: 2006 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (July), 423–430. https://doi.org/10.3115/1610075.1610135

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training, Research B*(February), 49. https://doi.org/ERIC #:ED108134

Klare, G. R. (1974). Assessing Readability. *Reading Research Quarterly*, *10*(1), 62. https://doi.org/10.2307/747086

Könsgen, R., Schaarschmidt, M., Ivens, S., & Munzel, A. (2018). Finding Meaning in Contradiction on Employee Review Sites — Effects of Discrepant Online Reviews on Job Application Intentions. *Journal of Interactive Marketing*, *43*, 165–177. https://doi.org/10.1016/j.intmar.2018.05.001

Korfiatis, N, Rodríguez, D., & Sicilia, M. . (2008). The impact of readability on the usefulness of online product reviews: a case study on an online bookstore. *World Summit on Knowledge Society*, 423–432.

Korfiatis, Nikolaos, García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, *11*(3), 205–217. https://doi.org/10.1016/j.elerap.2011.10.003

Kumar, N., Qiu, L., & Kumar, S. (2018). Exit, voice, and response on digital platforms: An

empirical investigation of online management response strategies. *Information Systems Research*, *29*(4), 849–870. https://doi.org/10.1287/ISRE.2017.0749

Lappas, T., Sabnis, G., & Valkanas, G. (2016). The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research*, *27*(4), 940–961. https://doi.org/10.1287/isre.2016.0674

Lee, Y. J., Besharat, A., Xie, K., & Tan, Y. (2018). Management Responses to Online Reviews: Helpful or Detrimental? In *ICIS 2017: Transforming Society with Digital Innovation*.

Li, C., Cui, G., & Peng, L. (2017). The signaling effect of management response in engaging customersA study of the hotel industry. *Tourism Management*, *62*, 42–53. https://doi.org/10.1016/j.tourman.2017.03.009

Li, C., Cui, G., & Peng, L. (2018). Tailoring management response to negative reviews: The effectiveness of accommodative versus defensive responses. *Computers in Human Behavior*, *84*, 272–284. https://doi.org/10.1016/j.chb.2018.03.009

Li, H., Zhang, Z., Janakiraman, R., & Meng, F. (2016). How review sentiment and readability affect online peer evaluation votes? An examination combining reviewer's social identity and social network. In *Proceedings of 2016 TTRA International Conference* (pp. 1–11). Retrieved from http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1224&context=ttra

Liang, S., & Li, H. (2019). Respond more to good targets: An empirical study of managerial response strategy in online travel websites. *E-Review of Tourism Research*, *16*(2–3), 215–223.

Liang, S., Schuckert, M., & Law, R. (2017). Multilevel Analysis of the Relationship Between Type of Travel, Online Ratings, and Management Response: Empirical Evidence from International Upscale Hotels. *Journal of Travel and Tourism Marketing*, *34*(2), 239–256. https://doi.org/10.1080/10548408.2016.1156613

Liu, J., Lin, C., Yunbo, C., Yalou, H., & Ming, Z. (2007). Low-Quality Product Review Detection in Opinion Summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 334–342).

Liu, X., & Law, R. (2019). Insights into managers' response behavior: Priority and effort. *International Journal of Hospitality Management*, *77*(August 2018), 468–470. https://doi.org/10.1016/j.ijhm.2018.08.010

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp

review fraud. *Management Science*, *62*(12), 3412–3427. https://doi.org/10.1287/mnsc.2015.2304

Lui, T. W., Bartosiak, M., Piccoli, G., & Sadhya, V. (2018). Online review response strategy and its effects on competitive performance. *Tourism Management*, *67*, 180–190. https://doi.org/10.1016/j.tourman.2018.01.014

Marx, P., & Nimmermann, F. (2018). Online Complaints in the Eye of the Beholder: Optimal Handling of Consumer Complaints on the Internet. In *ICIS 2017: Transforming Society with Digital Innovation* (pp. 0–9).

Mate, M. J., Trupp, A., & Pratt, S. (2019). Managing negative online accommodation reviews: Evidence from the cook islands. *Journal of Travel and Tourism Marketing*, *36*(5), 627–644. https://doi.org/10.1080/10548408.2019.1612823

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, *12*(8), 639–646. https://doi.org/10.1039/b105878a

Meng, F., Dipietro, R. B., Gerdes, J. H., Kline, S., & Avant, T. (2018). How hotel responses to negative online reviews affect customers' perception of hotel image and behavioral intent: An exploratory investigation. *Tourism Review International*, *22*(1), 23–39. https://doi.org/10.3727/154427218X15202734130422

Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*.

Mitchell, T. M. (1997). *Machine Learning*.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly: Management Information Systems*. https://doi.org/10.2307/20721420

Murphy, R. (2018). Local Consumer Review Survey 2018. *Bright Ideas / Research*.

Murphy, R. (2019). Local Consumer Review Survey 2019. *Bright Ideas / Research*.

Niu, R. H., & Fan, Y. (2018). An exploratory study of online review management in hospitality services. *Journal of Service Theory and Practice*, *28*(1), 79–98. https://doi.org/10.1108/JSTP-09-2016-0158

Olson, E. D., & Ro, H. (2019). Company Response to Negative Online Reviews: The Effects of Procedural Justice, Interactional Justice, and Social Presence. *Cornell Hospitality Quarterly*. https://doi.org/10.1177/1938965519892902

Pan, Y., & Zhang, J. Q. (2011). Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews. *Journal of Retailing*, *87*(4), 598–612. https://doi.org/10.1016/j.jretai.2011.05.002

Pavlou, P. A., & Dimoka, A. (2006). The nature and role of feedback text comments in online

marketplaces: Implications for trust building, price premiums and seller differentiation. *Information Systems Research*, *17*(4), 392–414. https://doi.org/10.1287/isre.1060.0106

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, *46*(1), 69–81. https://doi.org/10.1037/0022-3514.46.1.69

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*(C), 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Piehler, R., Schade, M., Hanisch, I., & Burmann, C. (2019). Reacting to negative online customer reviews: Effects of accommodative management responses on potential customers. *Journal of Service Theory and Practice*, *29*(4), 401–414. https://doi.org/10.1108/JSTP-10-2018-0227

Pikulski, J. J. (2002). Readability. *Houghton Mifflin*, 1–12.

Priante, A., Hiemstra, D., van den Broek, T., Saeed, A., Ehrenhard, M., & Need, A. (2016). #WhoAmI in 160 Characters? Classifying Social Identities Based on Twitter Profile Descriptions, 55–65. https://doi.org/10.18653/v1/w16-5608

Proserpio, D., & Zervas, G. (2017). Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews. *Marketing Science*, *36*(5), 645–665. https://doi.org/10.1287/mksc.2017.1043

Raju, A. (2019). Can reviewer reputation and webcare content affect perceived fairness? *Journal of Research in Interactive Marketing*, *13*(4), 464–476. https://doi.org/10.1108/JRIM-05-2018-0065

Roozen, I., & Raedts, M. (2018). The effects of online customer reviews and managerial responses on travelers' decision-making processes. *Journal of Hospitality Marketing and Management*, *27*(8), 973–996. https://doi.org/10.1080/19368623.2018.1488229

Rose, M., & Blodgett, J. G. (2016). Should Hotels Respond to Negative Online Reviews? *Cornell Hospitality Quarterly*, *57*(4), 396–410. https://doi.org/10.1177/1938965516632610

Salton, G., & Christopher, B. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Sandulescu, V., & Ester, M. (2015). Detecting Singleton Review Spammers Using Semantic Similarity. In *Proceedings of the 24th international conference on World Wide Web* (pp. 971–976).

Schroeder, M. A., Lander, J., & Levine-Silverman, S. (1990). Diagnosing and dealing with

multicollinearity. *Western Journal of Nursing Research*, *12*(2), 175–187. https://doi.org/10.1177/07399863870092005

Schuckert, M., Liang, S., Law, R., & Sun, W. (2019). How do domestic and international high-end hotel brands receive and manage customer feedback? *International Journal of Hospitality Management*, *77*(August 2018), 528–537. https://doi.org/10.1016/j.ijhm.2018.08.017

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Sheng, J. (2019). Being Active in Online Communications: Firm Responsiveness and Customer Engagement Behaviour. *Journal of Interactive Marketing*, *46*, 40–51. https://doi.org/10.1016/j.intmar.2018.11.004

Sheng, J., Amankwah-Amoah, J., Wang, X., & Khan, Z. (2019). Managerial Responses to Online Reviews: A Text Analytics Approach. *British Journal of Management*, *30*(2), 315–327. https://doi.org/10.1111/1467-8551.12329

Shin, H., Perdue, R. R., & Pandelaere, M. (2019). Managing Customer Reviews for Value Co-creation: An Empowerment Theory Perspective. *Journal of Travel Research*. https://doi.org/10.1177/0047287519867138

Smith, E. A., & Senter, R. J. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, 1–14.

Sparks, B. A., & Bradley, G. L. (2017). A "Triple A" Typology of Responding to Negative Consumer-Generated Online Reviews. *Journal of Hospitality and Tourism Research*, *41*(6), 719–745. https://doi.org/10.1177/1096348014538052

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

Sparks, B. A., So, K. K. F., & Bradley, G. L. (2016). Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management*, *53*, 74–85. https://doi.org/10.1016/j.tourman.2015.09.011

Sreejesh, S., & Anusree, A. (2016). The impacts of customers' observed severity and agreement on hotel booking intentions: moderating role of webcare and mediating role of trust in negative online reviews. *Tourism Review*, *71*(2), 77–89. https://doi.org/10.1108/TR-08-2015-0037

Sreejesh, S., Anusree, M. R., & Ponnam, A. (2019). Can online service recovery interventions benignly alter customers' negative review evaluations? Evidence from the hotel industry.

Journal of Hospitality Marketing and Management, *28*(6), 711–742. https://doi.org/10.1080/19368623.2019.1544958

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases, *185*(4157), 1124–1131.

van Noort, G., Willemsen, L. M., Kerkhof, P., & Verhoeven, J. W. M. (2014). Webcare as an Integrative Tool for Customer Care, Reputation Management, and Online Marketing. In *Integrated Communications in the Postmodern Era* (pp. 77–99). https://doi.org/10.1057/9781137388551.0008

Verhoeven, M. (2016). A brief introduction to QCA. *Methodische Zugänge Zur Erforschung von Medienstrukturen, Medienorganisationen Und Medienstrategien*, (November 2016), 173–196. https://doi.org/10.5771/9783845276885-173

Wallin Andreassen, T. (2000). Antecedents to satisfaction with service recovery. *European Journal of Marketing*, *34*(1/2), 156–175. https://doi.org/10.1108/03090560010306269

Weitzl, W., & Hutzinger, C. (2017). The effects of marketer- and advocate-initiated online service recovery responses on silent bystanders. *Journal of Business Research*, *80*(April), 164–175. https://doi.org/10.1016/j.jbusres.2017.04.020

Weitzl, W., Hutzinger, C., & Einwiller, S. (2018). An empirical study on how webcare mitigates complainants' failure attributions and negative word-of-mouth. *Computers in Human Behavior*, *89*(August), 316–327. https://doi.org/10.1016/j.chb.2018.07.012

Weitzl, W. J. (2019). Webcare's effect on constructive and vindictive complainants. *Journal of Product and Brand Management*, *28*(3), 330–347. https://doi.org/10.1108/JPBM-04-2018-1843

Weitzl, W. J., & Einwiller, S. A. (2019). Profiling (un-)committed online complainants: Their characteristics and post-webcare reactions. *Journal of Business Research*, (May), 1–14. https://doi.org/10.1016/j.jbusres.2019.05.035

Xie, K., Kwok, L., & Wang, W. (2017). Monetizing Managerial Responses on TripAdvisor: Performance Implications Across Hotel Classes. *Cornell Hospitality Quarterly*, *58*(3), 240–252. https://doi.org/10.1177/1938965516686109

Xie, K. L., & So, K. K. F. (2018). The Effects of Reviewer Expertise on Future Reputation, Popularity, and Financial Performance of Hotels: Insights from Data-Analytics. *Journal of Hospitality and Tourism Research*, *42*(8), 1187–1209. https://doi.org/10.1177/1096348017744016

Xie, K. L., So, K. K. F., & Wang, W. (2017). Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International*

Journal of Hospitality Management, *62*, 101–110. https://doi.org/10.1016/j.ijhm.2016.12.004

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

Xu, Y., Li, H., Law, R., & Zhang, Z. (2020). Can receiving managerial responses induce more user reviewing effort? A mixed method investigation in hotel industry. *Tourism Management*, *77*(November 2018), 103982. https://doi.org/10.1016/j.tourman.2019.103982

Xu, Y., Zhang, Z., Law, R., & Zhang, Z. (2019). Effects of online reviews and managerial responses from a review manipulation perspective. *Current Issues in Tourism*, 1–16. https://doi.org/10.1080/13683500.2019.1626814

Zhang, L., Gao, Y., & Zheng, X. (2020). Let's Talk About This in Public: Consumer Expectations for Online Review Response. *Cornell Hospitality Quarterly*, *61*(1), 68–83. https://doi.org/10.1177/1938965519864864

Zhang, X., Qiao, S., Yang, Y., & Zhang, Z. (2020). Exploring the impact of personalized management responses on tourists' satisfaction: A topic matching perspective. *Tourism Management*, *76*(June 2019), 103953. https://doi.org/10.1016/j.tourman.2019.103953

Zhang, Y., & Vásquez, C. (2014). Hotels' responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context and Media*, *6*, 54–64. https://doi.org/10.1016/j.dcm.2014.08.004

Zhang, Z., Li, H., Meng, F., & Li, Y. (2019). The effect of management response similarity on online hotel booking: Field evidence from Expedia. *International Journal of Contemporary Hospitality Management*, *31*(7), 2739–2758. https://doi.org/10.1108/IJCHM-09-2018-0740

Zwier, S. (2019). Webcare in healthcare: Providers responses to patients' online reviews. *British Journal of Health Care Management*, *25*(10), 1–7. https://doi.org/10.12968/bjhc.2018.0078

7 Appendices

Appendix A: Literature Matrix for Managerial Response Literature

Authors	Review Context/Industry	Data Source	Aspect of MR	Outcome Metrics	Main Findings
Casado-Díaz, A.B., Andreu, L., Beckmann, S.C., Miller, C. (2020)	Hotels	TripAdvisor, Twitter	MR style (accommodative, defensive, no action)	Booking intentions, brand attitude	1. No significant effect of MR on booking intentions was measured. 2. Accommodative response most strongly improves attitude. No response is worst.
Xu, Li, Law, Zhang (2020)	Hotels	Qunar (qunar.com)	Number of MRs	User reviewing effort	1. Users, who receive more MRs, put more effort into review writing. 2. If a hotel already has many high effort reviews, increased reviewing effort motivated by MR is attenuated. 3. Users who receive more MRs tend to allocate less effort, if many high effort reviews are already present for a hotel.
Zhang, L., Gao, Y., Zheng, X. (2020)	Hotels	Scenario based experiment	MR channel (private or public)		1. Reviewers tend to expect public responses to negative reviews. 2. Reviewers tend to care less about MR channel for positive reviews.
Zhang, X., Qiao, S., Yang, Y., Zhang, Z. (2020)	Hotels	TripAdvisor	Personalization of MR, Matching of MR and review topics	Rating	1. High levels of matching response leads to increase in rating.
Piehler, R., Schade, M., Hanisch, I., Burmann, C. (2019)	Hotels	Online experiment with fictitious reviews	Accommodative response strategy to negative reviews.	Purchase intention of potential customers	1. Both, explanation and compensation have a positive effect on purchase intention. 2. Explanation + Compensation is most effective.
Raju, A. (2019)	Restaurants	Experiment with fictitious reviews	MR content (specific vs. vague), MR source credibility	Perceived fairness	1. Specific MR content leads to higher perceived fairness, 2. High MR source credibility leads to higher perceived fairness.
Zwier, S. (2019)	Healthcare Physicians	National Health Service website (UK)	MR frequency MR quality	Comparison to non-medical provider MRs	1. Medical providers are less responsive to review content than non-medical providers. 2. Non-medical providers respond more to negative reviews.
Sreejesh, S., Anusree, M.R., Ponnam, A. (2019)	Hotels	Experiment with fictitious reviews	Response style after service failure	Attitude towards hotel, Purchase intention	1. Webcare can improve observing customers' attitude and purchase intention. 2. A mix of apology, prospective explanation and

					compensation is most effective.
Zhang, Z., Li, H., Meng, F., Li, Y. (2019)	Hotels	Expedia	MR number MR similarity	Hotel bookings	1. The number of MRs does not significantly influence the number of bookings for a hotel. 2. High similarity in MRs has a significant negative effect on number of bookings
Mate, M.J., Trupp, A., Pratt, S. (2019)	Hotels	TripAdvisor	MR Strategies used by managers	Perception of Brand & Reputation	A framework of strategies is developed. Including new aspects 'values', 'culture' and 'corrective statements'
Weitzl, W.J. (2019)	Online shopping & service	Online Survey & Online scenario experiment	MR Strategies (accommodative, defensive or no response) to constructive vs. vindictive complaints	Webcare satisfaction, brand satisfaction, brand image, brand loyalty, WOM intention,	1. Constructive complaints are best met with accommodative responses. Worst strategy is 'no response'. 2. Vindictive complaints are unaffected by all three strategies.
Sheng, J. (2019)	Hotels	A leading travel review site	MR Volume, speed & length	Review volume	1. Higher MR volume leads to higher review volume. 2. Higher MR speed leads to higher review volume. 3. Higher MR length leads to higher review volume.
Feng, W., Ren, W. (2019)	E-commerce (cosmetics & high tech digital products)	Taobao (taobao.com) & JingDong (jd.com)	Types of MRs, Similarities and differences in MRs to positive vs. negative reviews		1. Thanking, justification, promising and expectation of future purchase were the most frequent types of MR. 2. Different types of MR are used for positive vs. negative reviews. Positive: Thanking, advertisement, promising Negative: Justification, offer of solution
Sheng, J., Amankwah-Amoah, J., Wang, X., Khan, Z. (2019)	Hotels	A leading travel review site	MR vs. no MR, sentiment, length & speed	Rating	1. Ratings are higher if MRs are provided. 2. The effect is stronger for low satisfaction customers. 3. Longer responses tended to lead to higher increases in rating. 4. Response speed did not have a significant effect.
Alrawadieh, Z., Dincer, M.Z. (2019)	Hotels	TripAdvisor + Interviews with managers	MR rate, & content to negative reviews Occurrence of appreciation, apology, explanation, incentive in MRs.		1. Less than half of negative reviews received MRs. Many received standardized responses. 2. Managers emphasized the importance of answering to negative reviews. 3. Inconsistency between managers' account given and practice. 4. Appreciation is most frequent, then apology, then explanation, incentive almost never.

Weitzl, W.J., Einwiller, S.A. (2019)	Online shopping and service	Online survey + Scenario based online experiment	MR style (no response, defensive, accommodative) to 3 types of complainants	Negative word-of-mouth intention	1. The best strategy for constructive loyal customers is accommodative. 2. For constructive, unattached customers, all three are equally effective. 3. For vengeful loyal customers, all three are equally ineffective.
Schuckert, M., Liang, S., Law, R., Sun, W. (2019)	Hotels	TripAdvisor + Daodao (daodao.com)	MR frequency, quality (length)	Rating	1. Higher MR frequency led to higher ratings. 2. Responding to negative reviews is more effective for improving ratings. 3. Higher MR length to negative reviews leads to higher ratings.
Liu, X., Law, R. (2019)	Hotels	Qunar (qunar.com)	MR priority (Which reviews do managers respond to?) MR effort (length of MRs to different types of reviews)		1. Managers prioritize 'selected' reviews, negative reviews and long reviews. 2. Managers also exert more effort to those reviews.
Chen, W., Gu, B., Ye, Q., Zhu, K.X. (2019)	Hotels	Ctrip.com eLong.com	MR vs. no MR MR target, MR style,	Review volume, Rating	1. MRs increase review volume. 2. Rating change from MRs was not significant. 3. Managers should respond in detail to negative reviews and briefly to positive reviews.
Olson, E.D., Ro, H. (2019)	Restaurants & Hotels	Online questionnaire, scenarios based on TripAdvisor reviews and MRs	MR content (procedural justice, interactional justice, social presence)	Trust & purchase intention	1. Procedural justice and interactional justice in MRs lead to an increase in trust. 2. Social presence in MRs had no significant effect on trust.
Shin, H., Perdue, R.R., Pandelaere, M. (2019)	Hotels	Scenario based experiments	MR rate MR personalization	Customer empowerment Intention to co-create	1. Providing personalized MRs leads to more customer empowerment than not responding. 2. A personalized MR leads to more empowerment than an impersonal response. This effect is stronger in negative reviews. 3. These effects apply to posters of reviews and to readers.
Liang, S., Li, H. (2019)	Hotels	TripAdvisor	MR target	Review Volume, Rating	1. Responding to users who have posted multiple negative reviews, can boost review volume and rating.
Xu, Y., Zhang, Z., Law, R., Zhang, Z. (2019)	Hotels	TripAdvisor Expedia	MR rate	Hotel bookings,	1. Increasing MRs (on TripAdvisor) can reduce hotel bookings. 2. A lack of purchase verification on a website, leads to distrust.
Weitzl, W., Hutzinger, C., Einwiller, S. (2018)	E-commerce	Online questionnaire	MR style (no response, defensive, accommodative) + Antecedents (prior	Failure attribution, satisfaction, negative WOM	1. For few prior service failures, accommodative MR was most effective in decreasing failure attribution,

			failures, advocate webcare)		then defensive then no response. 2. For multiple prior service failures, effectiveness of the three styles was mostly similar.
Roozen, I., Raedts, M. (2018)	Hotels	Experiment with fictitious reviews + MRs	MR personalization	Booking intentions, WOM intentions	1. Personalized MRs have a stronger effect than general MRs. 2. This effect is stronger if reviews contain both positive and negative attributes.
Xie, K.L., So, K.K.F. (2018)	Hotels	TripAdvisor	MR targets (reviewer expertise)	Financial performance	1. MRs to users with high expertise (membership time, reviewer badge) have stronger positive effects than MRs to users with low expertise. 2. This effect did not appear for 'helpful votes'.
Chevalier, J.A., Dover, Y., Mayzlin, D. (2018)	Hotels	TripAdvisor, Expedia, Hotels.com	MR frequency (comparison of platforms with and without MR feature)	Review volume, review effort (length), Rating	1. MRs increase review volume. 2. MRs increase reviewing effort (as measured by length) 3. MRs lead to decrease in rating. 4. Responding mainly to negative reviews may encourage more (detailed) negative reviews.
Lui, T.-W., Bartosiak, M., Piccoli, G., Sadhya, V. (2018)	Hotels	TripAdvisor	MR frequency MR strategy (No response, respond to extreme reviews, respond to all)	Firm performance	1. MRs frequency is positively related to firm performance. 2. This effect is stronger for negative reviews. 3. Responding to extreme reviews has a stronger effect than responding to all reviews.
Könsgen, R., Schaarschmidt, M., Ivens, S., Munzel, A. (2018)	Employee reviews	Kununu.com + experiment with fictitious reviews and MRs	MR vs. non-MR	Trustworthiness, intention to pursue/avoid employment	1. MRs positively affect trustworthiness. 2. MRs lead to increased intention to pursue employment. 3. MRs lead to decreased intention to avoid employment.
Li, C., Cui, G., Peng, L. (2018)	Hotels, Online shopping	TripAdvisor + Experiment	MR style (accommodative, defensive), MR tailoring (product failure vs. ordinary negative review)	Hotel revenue, purchase intention	1. Tailored MRs lead to improved hotel revenues. 2. Tailored MRs enhance purchase intention. (product failure -> accommodative, ordinary negative review -> defensive
Lee, Y.-J., Besharat, A., Xie, K., Tan, Y. (2018)	Hotels	TripAdvisor	MR frequency, MR targeting (based on a hotel's review volume and valence)	Revenue	1. When valence and volume is low, MRs decreased revenue. 2. When valence and volume is high, MRs decreased revenue. 3. When one was high and the other was low, MRs increased revenue.

Marx, P., Nimmermann, F. (2018)	Online shopping	Experiment	MR type	Reader attitudes & behavior	1. MR type matters for attitudes more than behavior. 2. Combination of apology & redress evokes the most positive response.
Niu, R.H., Fan, Y. (2018)	Hospitality	Interviews	Presence and aspects of systematic MR strategy from a structure and process perspective		Structure includes: Formality, centralization, specialization Process includes: Review analytics, response customization, Integration
Meng, F., Dipietro, R.B., Gerdes, J.H., Kline, S., Avant, T. (2018)	Hotels	Online questionnaire with fictitious MRs	MR type (no response, negative response, service recovery response)	Image, Attitude, Intent to stay	1. Service recovery response most favorably affected image, attitude & intent to stay. 2. Negative response was more favorable than no response for image and attitude, but equally favorable for intent to stay.
Kumar, N., Qiu, L., Kumar, S. (2018)	Restaurants	Yelp.com	MR vs. non-MR, MR targeting	Firm performance, Performance of other local businesses,	1. Posting MRs increases firm performance. 2. Posting MRs decreases firm performance of other local businesses. 3. Managers tend to respond more to negative reviews.
Weitzl, W., Hutzinger, C. (2017)	Hospitality (coffee house)	Scenario based experiments + questionnaire	MR type (no response, accommodative, defensive)	Un/favorable brand-related reactions (brand attitude, trust, purchase intention, WOM intention, risk, failure attribution)	1. Accommodative response produces most favorable reactions. 2. No response and defensive response produce similarly favorable reactions.
Li, C., Cui, G., Peng, L. (2017)	Hotels	TripAdvisor	MR frequency, speed, length	Review volume, rating, helpfulness votes, popularity ranking	1. Higher MR frequency increases review volume and rating and popularity ranking, but not helpfulness votes. 2. Higher MR speed increases volume, rating, helpfulness votes and popularity ranking. 3. Higher MR length does not increase any of the metrics.
Proserpio, D., Zervas, G. (2017)	Hotels	TripAdvisor, Expedia	MR vs. no MR	Review volume, Rating, Review length	1. MRs increase rating (0.12 star) 2. MRs increase review volume (12%) 3. MRs lead to fewer but longer negative reviews.
Sparks, B.A., Bradley, G.L. (2017)	Hotels	TripAdvisor	What MR types (Acknowledgement, account, action) are used,		Propositions are developed 1. MRs to negative reviews increase brand perception. 2. Evaluations are more favorable if acknowledgement, account and action are present.

Xie, K., Kwok, L., Wang, W. (2017)	Hotels	TripAdvisor	MR number, speed, length, match rate,	Financial performance	1. Number of MRs increased revenue. 2. MR speed increased revenue. 3. MR length decreased revenue. 4. Match rate had no significant effect.
Xie, K.L., So, K.K.F., Wang, W. (2017)	Hotels	TripAdvisor	MR speed, length, repetition of topics, MR volume	Financial performance	1. Higher MR speed increases revenue. 2. Higher MR length increases revenue. 3. Higher MR volume increases revenue. 4. Higher repetition of topics decreases revenue.
Liang, S., Schuckert, M., Law, R. (2017)	Hotels	TripAdvisor, Ctrip	MR frequency, MR targets	Customer satisfaction	1. MR frequency increases customer satisfaction. 2. This effect is less severe for reviewers with multiple reviews.
Ho, V. (2017)	Hotels	TripAdvisor	MR types are extracted from MRs on TripAdvisor. Obligatory and optional 'moves' are identified		Obligatory: Acknowledging Problem, Expressing Feeling, Thanking Reviewer Optional: Continuing Relationship, Denying Problem, Greeting, Recognizing Reviewer's Value, Self Promoting
Rose, M., Blodgett, J.G. (2016)	Hotels	Scenario based experiment + TripAdvisor	MR vs. no MR MR type (apology with corrective action vs. apology with assurance of future satisfaction)	Company reputation	1. MRs to negative reviews increase company reputation. 2. When problems are perceived to be controllable, MRs have a more favorable impact. 3. 'Apology with corrective action' and 'apology with assurance of future satisfaction' are equally effective.
Sreejesh, S., Anusree, A. (2016)	Hotels	Scenario based experiment	MR vs. no MR	Booking intentions	In situations of high failure severity and high agreement among reviewers, MRs increase booking intentions.
Sparks, B.A., So, K.K.F., Bradley, G.L. (2016)	Hotels	Scenario based experiment	MR vs. no MR MR Source, MR Speed, MR Action frame, MR Tone of voice	Customer concern (attentive, caring, responsive) Trust inferences	1. Providing MRs increases perceived customer concern and trust inferences. 2. Higher MR speed increased favorable customer inferences. 3. More conversational 'human' voice, (vs. professional voice) increased favorable customer inferences. 4. Source and action frame did not produce significant results.
Kim, S.J., Wang, R.J.-H., Maslowska, E., Malthouse, E.C. (2016)	E-commerce	Scenario based experiment	MR vs. no MR (apology)	Behavioral intentions	MRs improved behavioral intentions of review viewers, but not of review posters.

Xie, K.L., Zhang, Z., Zhang, Z., Singh, A., Lee, S.K. (2016)	Hotels	TripAdvisor	MR frequency	Rating, Review volume, Revenue,	1. MRs increase review rating. 2. MRs increase review volume. 3. MRs positively moderate the relationship between review volume/valence and revenue.

Appendix B: Methodology for Managerial Response Literature Review

In order to find relevant literature for the domain of managerial responses, I developed a search string including multiple iterations of the managerial response term such as 'corporate response' or 'organizational response'. Additionally, the term 'webcare' was included, as it is sometimes used in the literature. I made use of the proximity operator W/1, which allows for results where the first word occurs within a one-word range of the second (scopus.com). This allows for expressions like 'response management' or 'corporate complaint response', casting a wider net to find relevant articles. Furthermore, I specified multiple sources of and terms for customer feedback. This was the search string I used in the Scopus database:

TITLE-ABS-KEY (((managerial W/1 response) OR (organizational W/1 response) OR (management W/1 response) OR (company W/1 response) OR (corporate W/1 response) OR (webcare)) AND ("online review" OR "e-WOM" OR "word of mouth" OR "customer review" OR "consumer review" OR "customer feedback" OR "consumer feedback"))

The search yielded 100 articles. Subsequently, I limited the search results to only include publications from the year 2016 and after, in order to obtain recent and relevant studies, resulting in 68 documents. I decided to stick to this timeframe since the online landscape is a quickly evolving space and it makes sense to review academic literature, in which circumstances resemble those of today[18]. Moreover, I deem the number and the content of relevant articles I found in this period sufficient, in order to form a balanced overview of the managerial response literature. After reviewing the 68 search results, I narrowed it down to 51 relevant articles. Out of these 51 articles, two had to be omitted because I was unable to obtain them, leaving 49 articles to be used in this literature review.

[18] This is not to imply that older studies are less valid, in fact, I refer to offline literature in various parts of the study. This merely refers to the search for related previous studies.

Appendix C: Codebook for Review Annotation

Codes

1 = The review text has high diagnosticity
0 = The review text has low diagnosticity

General Rules

1. Reviews are annotated purely based on the text of the review. Other features such as number of useful votes and rating score are not taken into account.

2. Diagnosticity is assessed based on content of the review text, rather than linguistic level or correctness. As long as the review text is intelligible, spelling or grammatical mistakes do not influence the decision.

3. Non-English reviews are excluded. Even though English reviews have been scraped, there may be a few non-English reviews if reviewers have indicated their language incorrectly. These reviews will be excluded from the annotated dataset.

Diagnosticity Explanations and Examples

Diagnosticity	**Review Feature**	**Explanation**	**Example Review Texts** (some have been shortened for display)
1 = High Diagnosticity	Service Feature	Describes/evaluates several service features	*"... The support, however, is friendly, but more importantly, the answers are quick and accurate. They care about the customer having a great experience. The panel is quick and easy to use. Furthermore, it does not break the bank. They offer a lot of payment options, so that is no concern as well. So far, I have used 2 of their services ... Everything was up immediately. The server was stable and all of us had low ms. The second service was a VPS, and again, it was up immediately, was stable and quick. ..."*
		Describes/evaluates one service feature	*"Good Support I first had some problems with my mc ftb revelation server, wrote the support a message, explaining everything etc. (they answered within 10 min) they migrated my server ..."*

	Experience with Company	Describes/evaluates experience with the company	*"...I once had a problem and within a short period of time the support helped me. However the website takes an oddly long amount of time to load and it's sometimes a bit hard to find things on it."*
	Call to action	Recommends and justifies a course of action to company or reader	*"Not able to make calls out . Every time a call is made it ask for Phone number and pin and when the phone number is entered it always says incorrect ... Please review your process as it is very inconvenient and redundant. Thanks"*
			"I am amazed how clear the voices come through. No lagging or anything. It's better than my landline. I would not hesitate to recommend this to anyone."
0 = Low Diagnosticity	Unspecific	Mentions, but does not justify, explain or describe any review features	*"Best VPS, Best support and they have best price"*
			"Rude staff and contributors. unskilled and unequipped."
	Uninformative	Contains no service features or experience with the company	*"So far I am very pleased"*
	Unintelligible	Review text is not intelligible	*"Kulli Kullirkfkfuvu"*

Appendix D: Regression Results without Review Volume

Table 10 Regression Results for Review Length (log), Review Volume Omitted

Variable	Base Model			MR Model		
	b	s.e.	p	b	s.e.	p
MRFreq				-0.2105	0.021	0.000
MRLength				0.0015	0.000	0.000
MRSpeed				-0.0256	0.004	0.000
MRRead	0.0139	0.001	0.000	0.0072	0.001	0.000
Rating	-0.3815	0.013	0.000	-0.3710	0.012	0.000
Status	0.0904	0.022	0.000	0.0330	0.021	0.120
Verified	-0.4166	0.021	0.000	-0.4956	0.021	0.000
UserReviews	-0.1522	0.019	0.000	-0.1335	0.018	0.000
Constant	5.3473	0.064	0.000	5.468	0.064	0.000
Adj. R squared	0.326			0.382		

Table 11 Regression Results for Useful Votes, Review Volume Omitted

Variable	Base Model			MR Model		
	b	s.e.	p	b	s.e.	p
MRFreq				-1.6216	0.097	0.000
MRLength				-0.0015	0.001	0.017
MRSpeed				-0.0128	0.020	0.519
MRRead	-0.0001	0.005	0.981	-0.0179	0.005	0.000
Rating	-1.6519	0.059	0.000	-1.6310	0.056	0.000
Status	-0.8154	0.099	0.000	-0.9497	0.097	0.000
Verified	-1.0275	0.097	0.000	-1.5107	0.099	0.000
UserReviews	-0.1902	0.087	0.000	-0.0735	0.084	0.383
Constant	3.0250	0.294	0.000	4.0660	0.292	0.000
Adj. R squared	0.300			0.351		

Table 12 Regression Results for Readability, Review Volume Omitted

Variable	Base Model			MR Model		
	b	s.e.	p	b	s.e.	p
MRFreq				0.2001	0.080	0.012
MRLength				0.0047	0.001	0.000
MRSpeed				-0.0537	0.016	0.001
MRRead	0.0176	0.004	0.000	0.0082	0.004	0.045
Rating	-0.5810	0.047	0.000	-0.5637	0.046	0.000
Status	0.2536	0.080	0.001	0.1715	0.080	0.033
Verified	-0.2907	0.078	0.000	-0.2780	0.081	0.001
UserReviews	0.0037	0.070	0.958	-0.0032	0.069	0.964
Constant	11.2571	0.236	0.000	11.0821	0.240	0.000
Adj. R squared	0.051			0.074		

Table 13 Regression Results for Diagnosticity, Review Volume Omitted

Variable	Base Model			MR Model		
	b	s.e.	p	b	s.e.	p
MRFreq				-0.1221	0.009	0.000
MRLength				0.0005	0.000	0.000
MRSpeed				-0.0129	0.002	0.000
MRRead	0.0067	0.000	0.000	0.0038	0.000	0.000
Rating	-0.1445	0.006	0.000	-0.1402	0.005	0.000
Status	0.0575	0.010	0.000	0.0307	0.009	0.001
Verified	-0.1747	0.009	0.000	-0.2142	0.009	0.000
UserReviews	-0.0981	0.008	0.000	-0.0872	0.008	0.000
Constant	1.2671	0.028	0.000	1.3466	0.028	0.000
Adj. R squared	0.292			0.353		