Artificial Intelligence for Online Review Platforms

Data Understanding, Enhanced Approaches and Explanations in Recommender Systems and Aspect-based Sentiment Analysis

Dissertation
zur Erlangung des Grades eines
Doktors der Wirtschaftswissenschaft
(Dr. rer. pol.)

eingereicht an der
Fakultät für Wirtschaftswissenschaften der
Universität Regensburg

vorgelegt von
Marcus Hopf (M.Sc. Mathematik)

Berichterstatter:
Prof. Dr. Bernd Heinrich
Prof. Dr. Günther Pernul

Regensburg, Mai 2022

Tag der Disputation: 07.07.2022

To my parents Karin and Ludwig.

Acknowledgements

First and foremost, I would like to thank Bernd Heinrich and express my deep gratitude to him for his outstanding supervision of my dissertation and his excellent guidance for my development as a young researcher over the last years. In particular, I would like to thank him for the great opportunity he gave to become a PhD student at his chair, his tremendous and continual support, his constructive and insightful feedback, the always thriving conversations, his thought-provoking impulses and his innovative ideas.

Second, I would like to thank Günther Pernul for his superb supervision of my dissertation and his precious and inspiring advice and ideas.

In addition, I would like to thank all other people who have supported, accompanied, guided and inspired me over the last years during my PhD, as this dissertation has been written with the direct and indirect support of many individuals, to whom I am very grateful. In particular, I would like to thank my co-authors and express deep gratitude to Alexander Schiller, Markus Binder, Michael Szubartowicz, Daniel Lohninger and Theresa Hollnberger for their superb collaboration, all the great and inspiring conversations, their huge and endless efforts, and the inspiring and pleasant work atmosphere full of mutual support. Moreover, I would like to give special thanks to Mathias Klier, Andreas Obermeier, Maximilian Förster, Michael Bortlik, Dominik Schön, Armin Steinwender, Heike Gorski and Leonie Kreuser. I also would like to thank all my other colleagues from University of Regensburg for their support and the fruitful conversations as well as all students who worked with me and supported my research. In addition, I would like to thank all my friends for accompanying me on my way to this dissertation. I am very grateful for sharing moments with all these great people who have enriched my live, guided me towards my aims and made me the person I am now.

Finally, I would like to express my heartfelt gratitude towards my family. In particular, I would like to thank Karin, Ludwig and Andrea for their unconditional love and their ever-present support they have always provided to me and, in particular, during the work on this dissertation. For this, I am deeply grateful.

Thank you!

Marcus Hopf May 2022

Contents

Remark: To facilitate selective reading, each paper in the dissertation is treated as its own manuscript with respect to abbreviations, figures, tables as well as general numbering and references and is thus self-contained.

1 Introduction

"As for the future, your task is not to foresee it, but to enable it."
Antoine de Saint-Exupéry (*1900; †1944)

1.1 Enabling the Future

The epoch-making and ever faster technological progress provokes disruptive changes, poses pivotal challenges but offers tremendous potential at the same time. Currently, artificial intelligence (AI) is "viewed as the most important disruptive new technology" (Benbya et al., 2020) raising new significant challenges to large organizations (Benbya et al., 2021). Given this reality with those recent technological advances, the vision of this dissertation is to contribute to enabling the future by unveiling the potential of AI for online review platforms and the consumers and businesses on these platforms in electronic commerce.

1.2 The Rise of AI

AI comprises theory, methods and techniques that help machines to analyze, simulate, exploit and explore human thinking processes and behavior (Lu, 2019). Amongst others, AI comprises fields such as big data, machine learning, artificial neural networks, deep learning, image and speech recognition, national language processing (NLP) and predictive analysis (Lu, 2019; Russel and Norvig, 2016). Furthermore, AI offers continuous innovation for the improvement and potential replacement of human tasks and activities with a wide range of applications in practice, such as finance, healthcare, manufacturing, retail, supply chain, logistics and utilities (Milana and Ashta, 2021), and is also subject to many research disciplines, such as economics, healthcare, information systems, computer science, neuroscience, psychology, philosophy and linguistics (Russel and Norvig, 2016). Especially for businesses, AI is already used for improved decision making based on large amounts of data (Janssen et al., 2017). For example, recommender systems (RS) enable to generate personalized recommendations and advertising for individual consumers with a plethora of alternative items (e.g., products or services) available (Li, 2019), while NLP applications such as sentiment analysis allow for feasible analyses and extractions of information from millions of textual documents (Li et al., 2018a). Here, the key factor for success of AI applications is the recent advent of technological progress. This allows to utilize data strategically for deriving insights and knowledge by processing and leveraging large amounts of data and thus, data becomes a continuously changing asset able to unleash new revenue opportunities for monetization (Firouzi et al., 2022). In particular, the ongoing rapid advances in the fields of memory and computation technologies as well as in information systems make AI very powerful. On the one hand, it is necessary to store large volumes of data, which serve as the basis of AI to operate on. On the other hand, new information system technologies such as social media (e.g., Yelp, Twitter, Facebook, LinkedIn or WhatsApp) and sensors of connected smart devices (e.g., smart IoT-devices or smart phones) allow to access vast amounts of data that capture valuable real-world information of high variety by high velocity (Baesens et al., 2016). For example, personalized recommendations to consumers can be made by means of analyzing the preferences

of consumers from large amounts of multi-faceted consumer data that are generated day-by-day (Ricci et al., 2011). Together, the advances in memory technology and information systems give rise to big data (Heinrich and Hristova, 2014; Sänger et al., 2014), which is characterized by the three-dimensional increase of data in volume, velocity and variety, called the "three V's" (Abbasi et al., 2016; Mauro et al., 2015). In addition, the progress in computation technology then allows to take full advantage of big data in an affordable way and thus, enables the adoption of AI on a large and broad scale in many fields and applications.

The resulting high relevance of AI is further indicated by the following figures. The Stanford annual AI index report states that the total global investment in AI in 2020 was $68 billion, which is an increase by 40% relative to 2019 (Zhang et al., 2021). Further, companies leveraging AI for their services generate large revenues. For example, Google's online ads business, which utilizes AI for personalized advertising, achieved a revenue of $147 billion in 2020 constituting over 80% of the company's total revenue (Graham and Elias, 2021). Moreover, the global software market focusing on AI is forecast to grow rapidly with an estimated increase by 21.3% to $62.5 billion in 2022 relative to 2021 and to reach up to $126 billion by 2025 (Rimol, 2021; Statista, 2020). Besides the high relevancy in practical applications, also AI's relevancy in research vastly increases, as indicated by the growing number of publications regarding AI. In particular, the annual growth rate of number of AI journal publications strongly increased from 19.6% in 2019 to 34.5% regarding 75 thousand AI journal publications in 2020 (Zhang et al., 2021).

1.3 The Role of Online Reviews Platforms

A broadly discussed and highly relevant field that already shows high AI adaption and still offers tremendous potential for future AI development is electronic commerce (Bawack et al., 2022; Song et al., 2019). With the emergence and proliferation of the internet, electronic commerce has developed into a major disruptor for traditional commerce and retail. Indeed, in 2021, worldwide retail electronic commerce sales reached $4.9 trillion, with its share of global retail estimated to rise up to 24% in 2025 (Cramer-Flood, 2022). Hereby, online review portals, as for example Yelp, Amazon, Tripadvisor and Google Maps, play a very important role in electronic commerce. While these portals assist consumers to find relevant items day-by-day, they also act as reputation systems (Sänger and Pernul, 2018). Here, they provide millions of user-generated online consumer reviews, which constitute electronic word-of-mouth (Jabr et al., 2020) and thus, are a major purchase influence factor (Liu et al., 2019; Yi et al., 2019) and assist consumers at making better selection decisions. Online consumer reviews comprise rich information and typically consist of a star rating (e.g., one to five stars) representing the overall consumer assessment of an item and a textual part with fine-grained consumer assessments (Sun et al., 2019; Yin et al., 2014). This emphasizes that online consumer reviews are important instruments for consumers to overcome information asymmetries about items (Feng et al., 2019) and in addition, they also constitute important performance indicators for businesses and online review platforms (Jabr et al., 2020). Besides this high relevancy and strong impact of online review platforms in practical business applications, these platforms and its provided online consumer reviews are major and highly attractive research topics in the field of information systems and, in particular, in the subfields RS and sentiment analysis (Baum and Spann, 2014; Jabr et al., 2020; Sun et al., 2019).

With the rising popularity of online review platforms, new challenges for consumers, businesses and the platforms themselves emerge that offers large potentials and new opportunities. One of the biggest challenges here is the massive amount of information available for consumers, such as the provided items and online consumer reviews. For example, it was estimated that Amazon had already hosted around 250 million reviews on their platform in 2018 (Nguyen, 2018), while Amazon provided 353 million products on their marketplace in 2021 (Buck, 2022). Moreover, users of Google Maps add more than 20 million pieces of information to the platform every day (Galov, 2022). Furthermore, Tripadvisor provided more than one billion reviews on its platform on nearly eight million accommodations, restaurants, experiences, airlines and cruises (Kaufer, 2022). These large amounts of items and reviews pose the major problem of information overload to consumers, businesses and platforms (Hu and Krishen, 2019). Consumers face the information overload problem when they are purchasing or selecting items with millions of alternatives available (Ricci et al., 2011). Similarly, platforms have to filter items and reviews to show to consumers visiting their platforms, as it is not feasible to present consumers millions of items and reviews. Moreover, businesses face the information overload problem when they try to assess the consumer feedback (e.g., needs or criticism) contained in multiple thousands of reviews in order to gain new insights regarding their consumers, products and services (Siering and Janze, 2019). Here, also platforms strive to extract the fine-grained consumer sentiments from millions of reviews in order to better understand the consumers and their preferences. In particular, the analysis of millions of reviews is further aggravated by the problem that the fine-grained consumer sentiments contained in online consumer reviews are given in unstructured, textual form.

1.4 AI for Online Reviews Platforms

Here, the potential of AI comes into play, as it allows to process such high volume of data of high variety. On the one hand, RS have been designed to overcome the problem of information overload when selecting or filtering items or reviews with a high volume of alternative items available. Thus, RS assist at filtering relevant items and reviews for consumers and platforms (Ricci et al., 2011). RS are subdivided into the types collaborative RS, content-based RS and hybrid RS, which are all based on the same principle of filtering items that match the preferences of consumers. While collaborative RS utilize ratings of other, similar consumers item filtering, content-based RS instead utilize the characteristics of the items, which a consumer has liked in the past, to infer other similar items for recommendation. More precisely, collaborative RS focus on data that enables to infer consumer preferences (e.g., ratings) and content-based RS focus on item content data (e.g., item features) that presents the characteristics of items. To overcome limitations of collaborative and content-based RS, hybrid RS were created by combining both approaches. On the other hand, NLP and text analytics, such as sentiment analysis, have emerged for coping with large amounts of unstructured, textual data of high variety, for instance, to harness the fine-grained consumer assessments in millions of online consumer reviews (Müller et al., 2016). Here, aspect-based sentiment analysis (ABSA) is the field of NLP that strives for automated analysis of fine-grained consumer sentiments in texts regarding certain aspects of an item (Angelidis and Lapata, 2018; Binder et al., 2019; Pontiki et al., 2016; Xu et al., 2019).

The high relevancy of these challenges led to the rise of multiple AI approaches and applications of RS and ABSA in research and practice (Birjali et al., 2021; Jain et al., 2021; Jesse and Jannach, 2021; Karimi et al., 2018). In particular, RS help to improve consumer trust and business performance results (e.g., increased sales or consumer retention) (Jannach et al., 2019; Panniello et al., 2016). For example, 60% of the clicks on the home screen of YouTube, 35% of sales on Amazon and 75% of what people watch on Netflix are based on recommendations, which, for example, results in Netflix‘s estimated business value of recommendation and personalization of more than $1 billion per year (Jannach and Jugovac, 2019). While ABSA enables to enhance RS regarding their performance in predicting relevant items (Dang et al., 2021; Musto et al., 2017; Stumme and Hotho, 2013), aspect-based sentiments can also be used to explain recommendations on platforms (Yang et al., 2013), which can further increase business performance results. In addition, ABSA is highly valuable for businesses as the fine-grained sentiments of online consumer reviews are key for information systems and business decisions in electronic commerce for product development, services offerings and forecasting future demands (Shrestha et al., 2021; Siering and Janze, 2019).

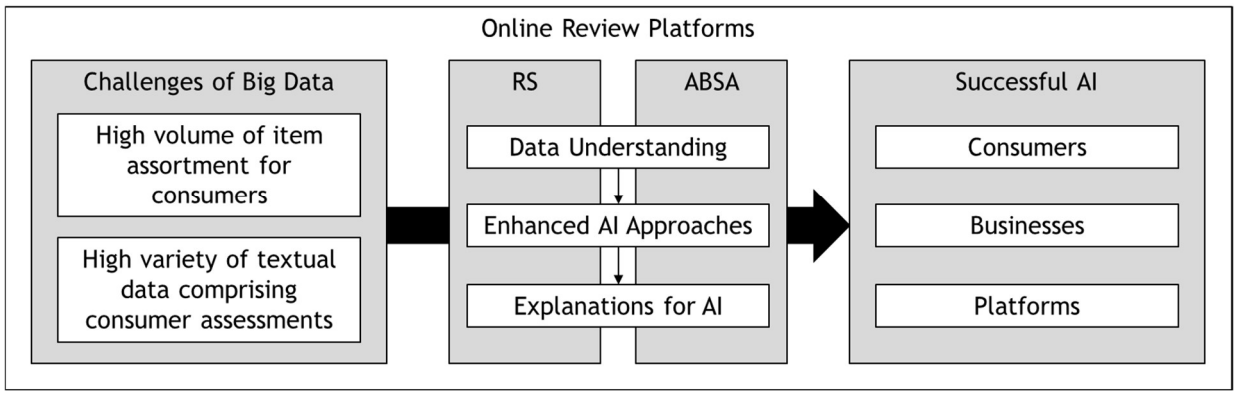

Figure 1-1 *The challenges of big data enable successful AI for online review platforms.*

Despite the fact that AI’s relevancy is forecast to increase in the next years, there still exists a large and manifold potential of AI that can be exploited to solve existing and future challenges and thus, to enable a fruitful and beneficial future for AI in general and especially in electronic commerce (Benbya et al., 2020; Jesse and Jannach, 2021; Lu, 2019). To enable successful AI for consumers, businesses and platforms on online review platforms, it is crucial to face and take advantage of the challenges of big data regarding the high volume of item assortment for consumers and regarding the high variety of textual data comprising consumer assessments. Therefore, it is promising to *understand* the data of online consumer reviews, to *enhance* existing AI approaches in RS and ABSA as well as to build trust in these AI approaches by means of *explanations* for their behavior and outcomes. An overview of this concept is given in Figure 1-1. First, data understanding of online consumer reviews is vital for successful applications of AI for online review platforms. Here, it is key to understand how consumers arrive at their overall rating of an item and, in particular, what factors (e.g., prior consumer ratings or contextual circumstances; e.g., cf. Richthammer and Pernul, 2020) influence these overall item assessments (Binder et al., 2019). While such insights from data understanding of online consumer reviews can be used for marketing campaigns and product development, they are particularly beneficial for improving AI approaches, which are based on that data. Second, enhancing existing approaches in RS and ABSA

to exploit the potential of big data of online consumer reviews is beneficial for improved performances of RS and ABSA. As high volume variety of data is available at online review platforms, existing RS approaches can be enhanced by integrating data of different sources (e.g., different platforms) and existing ABSA approaches can be enhanced by combining different types of data (e.g., numeric ratings and review texts). Third, explanations for RS and ABSA are essential for the success and adoption of AI at online review platforms, as, in general, most AI applications exhibit a black-box nature, which means that it is hardly possible to directly comprehend their predictions and outcomes without further assistance. Thus, explanations help to build trust and enable to justify business decisions based on AI (Bedué and Fritzsche, 2022; Kruse et al., 2019). In particular, the behavior of AI approaches can be explained by analyzing how they respond to different characteristics of input data (e.g., amount of missing values) as well as by replicating their outcomes with comprehensible surrogate models (e.g., reconstructions of its predictions). In total, the triad of data understanding, enhanced approaches and explanations of AI is key for a beneficial and fruitful future of AI for online review platforms. Therefore, the goal of the dissertation at hand is to extend the existing body of knowledge by giving major contributions to each part of this triad in the highly relevant field of AI for online review platforms.

1.4.1 Data Understanding of Online Consumer Reviews

Data understanding of online consumer reviews helps to comprehend the preferences, the item assessments and the needs of consumers. Thus, deriving insights from such data understanding is vital for marketing campaigns, product development and, in particular, for enhancing existing AI approaches (e.g., in RS or ABSA). As consumers assess items at different points in time and with reviews comprising an overall rating and a textual part, it is highly interesting to analyze temporal dynamics of consumer ratings as well as to disentangle these overall ratings by means of the provided fine-grained consumer sentiments in the review texts. Analyzing these two factors helps to build a thorough understanding of how consumers arrive at their overall rating of items.

Review ratings constitute the overall assessments of consumers regarding items. A central characteristic of these ratings is that they are generated at different points in time and in sequential order. Therefore, extant literature has already recognized the importance of sequential and temporal rating dynamics (e.g., Askalidis et al., 2017; Godes and Silva, 2012). However, these works have only shed light upon short-term dynamics. Existing findings, such as a decreasing rating trend in the short term caused by initial disconfirmation of consumers (Li and Hitt, 2008) thus cannot explain the long-term rating dynamics, which are particularly important as many items receive ratings over a long period of term. While it is interesting whether short-term dynamics continue in the long term, it is particularly relevant whether short-term dynamics exhibit any impact on subsequent ratings in the long term. In particular, strong initial rating dynamics resulting from too high ratings by early adopters (Hu et al., 2017; Li and Hitt, 2008) could show major impact on subsequent rating trends in the long term (e.g., an item's average rating). Here, extant literature has only analyzed effects of the very first rating on the trend of the subsequent ratings in the short term. Summing up, extant literature neither analyzes long-term rating dynamics (e.g., the rating trend continues in the long term) nor the impact of initial rating dynamics on the long-term

average rating. Therefore, the first research question of this dissertation addressed by the paper in Section 2.1 is as follows.

RQ1. *What are the long-term sequential and temporal dynamics in online consumer ratings and what is the impact of initial rating dynamics on the long-term average rating of items?*

To address this research question, regression models are applied to an extensive long-term review dataset in the domain of restaurants, spanning 13 years with more than 1.9 million online consumer reviews. The analysis yields two novel key findings extending the existing body of knowledge. First, a novel long-term sequentially increasing rating trend is found which leads to a U-shaped relationship between ratings and their order, indicating a second disconfirmation. Second, it is revealed that initial rating dynamics have significant impact on long-term average ratings, indicating a lasting effect based on initially stimulated expectations. Thus, "initially controversial" items achieve a lower long-term average rating. In particular, items initially receiving high ratings exhibit the lowest long-term average rating, emphasizing a strong lasting negative effect based on initially stimulated expectations.

Online consumer reviews comprise rich information relevant for consumers, businesses and platforms (Yin et al., 2014). Thus, a major goal of research is to derive insights about the relationship between review ratings, commonly given on a five-tier star rating scale, and review texts in order to understand why consumers rated items the way they did (Gutt et al., 2019). Therefore, it is important to leverage versatile features of different perspectives (e.g., the feature food quality of a restaurant from the perspective item aspects) derived from review texts to explain associated review star ratings. To enable comprehensible and well-founded insights regarding online consumer reviews, it is vital to base the analysis on easy-to-interpret features, which are features that can be traced back to its semantically related feature terms in the review texts. While extant literature has already conducted analysis using such features to explain associated ratings (e.g., Jabr et al., 2018; Xiang et al., 2015), these works have only considered selected features of at most two different feature perspectives in their analyses, enabling only a partial view. In addition, these works have thus not been able to investigate the contribution of multiple individual feature perspectives to the explanatory power of their proposed models. However, this would give important insights, such as that user characteristics (e.g., personality traits of consumers; cf. Goldberg, 1990) – a feature perspective rarely discussed in related literature – could show a key contribution to the explanation of review ratings, which would call for researchers to incorporate this feature perspective in their analysis of online consumer reviews. Therefore, the second research question of this dissertation addressed by the paper in Section 2.2 is as follows.

RQ2. *To what extent can features of different feature perspectives explain star ratings in online consumer reviews and how much does each individual feature perspective contribute to this explanatory power?*

To address this research question, the four feature perspectives (each comprising easy-to-interpret features) item characteristics, item aspects, user characteristics and user contexts are unified into one single model used for explaining star ratings. For the extraction of easy-to-interpret features for those perspectives from millions of review texts, Google's state-of-the-art deep learning language model BERT is applied (Devlin et al., 2019). The evaluation on three large real-world

datasets comprising experience goods (e.g., restaurants or movies) and search goods (e.g., laptops) shows that the proposed feature perspectives allow for explaining star ratings considerably well by means of 65-70% explanatory power. Here, all four feature perspectives show substantial contributions, with item aspects having the highest one. In addition, the rarely discussed feature perspective user characteristics yielded the second highest contribution.

1.4.2 Enhanced Approaches in RS and ABSA

Enhanced approaches in RS and ABSA allow for improved personalized recommendations and fine-grained sentiment extraction of higher quality and thus, enable the exploitation of the potential of AI for online review platforms. Moreover, the data understanding of online consumer reviews allows for targeted enhancements of RS and ABSA. Since RS operate on consumer ratings and item content data, the derived insights regarding rating dynamics as well as item aspects and item characteristics can be leveraged. For ABSA, the relationship between fine-grained sentiments towards feature perspectives and the overall review rating enables many opportunities. Thus, it is especially promising to improve existing RS approaches by integrating item content data of different platforms and to enhance existing ABSA approaches by combining overall review ratings and textual review data. Considering these two factors allows for improved personalized recommendations and high-quality sentiment extraction.

A big wave of researchers and practitioners enhancing and developing new algorithmic approaches for RS was initiated by the Netflix Grand Prize challenge in the years between 2006 and 2009, where matrix factorization approaches became very popular (Koren, 2009). Since then, many sophisticated approaches based on deep neural networks techniques (e.g., Xue et al., 2017; Zhang et al., 2018) have become dominant in the field, as they proclaimed substantial progress over the state-of-the-art (Dacrema et al., 2021). However, Dacrema et al. (2021) criticize that many of those approaches are either not reproducible or could be outperformed by already existing techniques such as matrix factorization or linear models. Thus, the field has arrived at a certain level of stagnation (Dacrema et al., 2021). However, as RS are algorithms that operate on data, increasing the quality of that data seems promising for alleviating that stagnation (Richthammer and Pernul, 2020; Sar Shalom et al., 2015). Here, existing literature supports that potential as they found that bad data quality impedes to derive insights from that data (Ghasemaghaei and Calic, 2019; Heinrich et al., 2018a; Vielberth et al., 2021) and reported that low data quality causes organizations losses of $15 million on average per year (Moore, 2018). In particular, the concept of big data comprising the “three V’s” is thus extended by a fourth “V” for veracity that represents the data quality for big data (Mauro et al., 2015). Especially for recommender systems, examining the item content data (i.e., features and feature values of items) and achieving a more complete view on these items seems to be promising, as “some representations capture only certain aspects of the content, but there are many others that would influence a user’s experience” (Picault et al., 2011). Hence, large potential to improve recommendations resides in increasing the data quality dimension completeness via extending item content data with additional data from other, similar data sources. Here, matching items in different datasets by means of duplicate detection seems highly promising (Heinrich et al., 2018b). However, extant literature (e.g., Li et al., 2018b; Ntoutsi

and Stefanidis, 2016) has not yet exploited that potential and therefore, the third research question of this dissertation addressed by the paper in Section 3.1 is as follows.

RQ3. *How can an item content dataset be systematically extended with respect to the data quality dimension completeness, aiming to improve recommendation quality?*

To address this research question, a procedure for the systematic extension of an item content dataset with an additional dataset and missing value imputation is proposed. The procedure is evaluated by means of two real-world datasets in the domains of restaurants and movies, which constitute commonly analyzed domains in research on information systems in electronic commerce. The results show that the presented procedure is indeed effective in enabling improved recommendations by increasing data quality.

Considering the large potential of the textual parts of online consumer review, fine-grained sentiments of consumers extracted by ABSA are key for business decisions in electronic commerce for product development, services offerings and forecasting future demands. In particular, the classification of fine-grained sentiments is vital, as it assigns a sentiment value (e.g., positive, neutral or negative) to a text segment of a review (e.g., a part of a sentence referring to one specific aspect of an item). The standard approach for fine-grained sentiment classification is training a supervised NLP classification model on labeled data comprising text segments with associated sentiment classes (Pannala et al., 2016). Here, the performance of supervised approaches highly depends on the amount and quality of labeled training data, which requires time-consuming and error-prone human labeling efforts. To mitigate this problem, novel techniques based on multiple instance learning (MIL) have been developed for economical fine-grained sentiment classification in online consumer reviews (Angelidis and Lapata, 2018; Kotzias et al., 2015; Pappas and Popescu-Belis, 2017). The basic idea of MIL is to infer fine-grained sentiments for review text segments (i.e., instances) from the associated overall review rating. Using MIL can be viewed as backpropagating the coarse-grained review-level sentiment information (i.e., the overall rating) onto the fine-grained sentiments of review text segments. Because of the high volume of available online consumer reviews with associated overall review ratings, MIL is well suited for economical fine-grained sentiment classification in online consumer reviews. Nevertheless, using supervised classification models is well established in the literature due to very good performances (Hoang et al., 2019; Pappas and Popescu-Belis, 2017). Given these complementary advantages of MIL and supervised models, it seems promising to combine both techniques to a MIL approach enhanced with partly supervision (MILPS) in order to improve the performance and efficacy of fine-grained sentiment classification. Yet, the extant literature in MIL for ABSA (cf. above) has not investigated such a combination and thus, the fourth research question of this dissertation addressed by the paper in Section 3.2 is as follows.

RQ4. *How can instance labels be incorporated into MIL for partly supervision, aiming at improved performance of fine-grained sentiment classification?*

To address this research question, a MILPS approach is proposed that extends MIL with partly fine-grained supervision by incorporating instance sentiment labels. Evaluating a state-of-the-art supervised model, a state-of-the-art MIL model and the MILPS model on a dataset comprising

online consumer reviews for restaurants yields that the MILPS model can enhance these state-of-the-art approaches in a very economical manner.

1.4.3 Explanations for RS and ABSA

Explanations for RS and ABSA are essential for the success and adoption of AI at online review platforms, as they are the key to build trust in existing, new and enhanced AI approaches and to comprehend and justify their behavior and outcomes. Since AI approaches heavily depend on the data they are based on as well as on their underlying model, it is reasonable to strive for explanations regarding those two components. On the one hand, AI approaches (e.g. enhanced RS) can be explained by analyzing how their outcomes depend on the input data. For example, in the case of enhanced RS approaches, it is interesting to investigate and explain why specific changes of input data characteristics (e.g., the amount of missing values) lead to improved recommendations. On the other hand, for many fields, such as electronic commerce or finance, it is highly promising to leverage modern NLP models for ABSA and use the derived sentiment insights for business decisions and applications. However, as these AI models exhibit a black-box nature, which means that it is hardly possible to trace back their predictions and outcomes without further assistance, explanations for such models (e.g., reconstructions of its predictions) are mandatory for their usage, deployment and a successful adoption.

As outlined in Section 1.4.2, large potential for enhancing RS (i.e., to improve the accuracy of RS's item relevancy predictions) resides in increasing the quality of data that RS are operating on (Richthammer and Pernul, 2020; Sar Shalom et al., 2015). Here, increasing the completeness via extending a dataset with features from another dataset and imputation of missing feature values is particularly promising, as such data quality improvement measures are feasible in many application scenarios. For example, in electronic commerce, there is an increasing number of heavily competing online review platforms with each of these platforms building and maintaining their own individual datasets. As the exploitation of external data sources requires valuable resources, it is vital for platforms to firstly examine the impact of increased completeness on the prediction accuracy of RS and whether other factors influence this impact. In particular, the increase in completeness affects items, consumers and features. Therefore, it is highly relevant to analyze what moderating effects the increased completeness of these three components has on the impact on recommendations. As extant literature (e.g., Ozsoy et al., 2015; Sar Shalom et al., 2015) has not yet analyzed the impact of completeness on RS or such effects, the fourth research question of this dissertation addressed by the paper in Section 4.1 is as follows.

RQ5. *Does the amount of available item features and the amount of filled up missing item feature values influence the prediction accuracy of recommender systems?*

To address this research question, a literature-based theoretical model is developed and hypotheses addressing the different aspects of this question regarding items, consumers and features are derived and substantiated. In particular, these hypotheses focus on the impact of adding features and of filling up missing feature values on the prediction accuracy of recommendations. The hypotheses are tested on two datasets from leading online review platforms in the domains of restaurants and movies. The results of these tests show that recommendations of RS are significantly more accurate when more features and feature values (i.e., completeness is increased)

are available. Furthermore, the results yield that the impact of completeness on prediction accuracy is moderated by the amount of increased completeness per items and per consumers. In addition, while adding features with many values leads to a higher increase in prediction accuracy, adding features with a high diversity does not lead to a higher increase in prediction accuracy compared to adding other features, which is contrary to extant literature (e.g., Mitra et al., 2002; Tabakhi and Moradi, 2015).

Utilizing extracted information and derived insights from ABSA for information systems and for decision making is promising for many fields such as electronic commerce or finance (Repke and Krestel, 2021; Shrestha et al., 2021). Here, language models such as BERT, which are deep learning AI models, constitute a major breakthrough and achieve leading-edge results in many applications of text analytics (e.g., ABSA for online consumer reviews). However, for a broad adoption of language models in research and practical applications, justifications and explanations for the outcomes and predictions of such black-box AI models are indispensable. Furthermore, regulations such as the General Data Protection Regulation of the European Union even impose an extensive right for explanations of automated data processing systems in general and thereby, lay the foundation to enforce algorithmic auditing in companies (Casey et al., 2019). In particular, as large amounts of ABSA predictions for newly-generated and hitherto unknown textual data are used on a daily basis in different information systems and for important business decisions, global explanations for language models are necessary. In contrast, a local explanation for each single prediction would require huge efforts for manual checking and thus, global explanations are required. Amongst different alternatives, rule-based explainable AI (XAI) approaches are recognized as a promising way to explain AI models, as they preserve the AI model itself and thus, its high performance, while offering post-hoc reconstructions as explanations (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020). In particular, as ABSA focuses on processing texts of natural language, global reconstructions based on linguistic information (so-called linguistic rules) seem promising. For assessing the quality of global reconstruction, its fidelity and comprehensibility have to be measured (Guidotti et al., 2019). As the comprehensibility of rule-based reconstructions is adaptable (e.g., by varying rule length), it is further interesting to analyze how different levels of comprehensibility affect fidelity. As extant literature (e.g., Ribeiro et al., 2018; Szczepański et al., 2021) lacks approaches for global reconstructions of the predictions of language models, the sixth research question of this dissertation addressed by the paper in Section 4.2 is as follows.

RQ6. *How can language model predictions be globally reconstructed by means of linguistic rules to balance fidelity and comprehensibility of the global reconstruction?*

To address this research question, a global XAI approach based on linguistic rules for reconstructing predictions of language models is proposed and the trade-off between fidelity and comprehensibility is analyzed. For evaluation, the language model BERT for aspect term and sentiment term detection in two datasets of the domains laptops and restaurants is analyzed. The results show that the proposed XAI approach based on linguistic rules is suited for a global reconstruction of BERT's predictions in online consumer reviews and, in particular, allows for balanced setups with respect to the trade-off between comprehensibility and fidelity of the reconstruction.

1.5 Structure of Dissertation

The dissertation comprises six research papers that address the research questions raised in Section 1.4. An overview of these papers grouped by their research goals and providing information regarding the papers' authors, institution of review and publication, current status and positioning in the dissertation is given by Table 1-1.

The remainder of the dissertation is structured as follows. The Sections 2, 3 and 4 contain the six research papers arranged within the research goals *data understanding of online consumer reviews*, *enhanced approaches in RS and ABSA* and *explanations for RS and ABSA* for online review platforms. Section 5 concludes the dissertation by presenting its major findings, summarizing its implications for research and practice and outlining remaining opportunities that could be starting points for future works regarding AI in electronic commerce and beyond.

Research Goal	Chapter and Title	Authors	Institution of Review and/or Publication	Current Status
Data Understanding of Online Consumer Reviews	Chapter 2.1: *Long-term Sequential and Temporal Dynamics in Online Consumer Ratings*	Bernd Heinrich Theresa Hollnberger Marcus Hopf Alexander Schiller.	European Conference on Information Systems (ECIS)	This paper is **accepted for publication** in *Proceedings of the 30th European Conference on Information Systems*.
	Chapter 2.2: *The Way to the Stars: Explaining Star Ratings in Online Consumer Reviews*	Markus Binder Bernd Heinrich Marcus Hopf Michael Szubartowicz	Decision Support Systems (DSS)	This paper is **under review in revision** for publication in the journal *Decision Support Systems*.
Enhanced Approaches in RS and ABSA	Chapter 3.1: *Something's Missing? A Procedure for Extending Item Content Data Sets in the Context of Recommender Systems*	Bernd Heinrich Marcus Hopf Daniel Lohninger Alexander Schiller Michael Szubartowicz	Information Systems Frontiers (ISF)	This paper is **accepted** and **published** in Volume 24, Issue 1 in the journal *Information Systems Frontiers*.
	Chapter 3.2: *Leveraging Fine-grained Supervision to Improve Multiple Instance Learning for Fine-grained Sentiment Classification in Online Consumer Reviews*	Marcus Hopf	European Conference on Information Systems (ECIS)	This paper was **under review** at *30th European Conference on Information Systems*. After rework, it will be submitted to *Transactions of the Association for Computational Linguistics.*
Explanations for RS and ABSA	Chapter 4.1: *Data Quality in Recommender Systems: The Impact of Completeness of Item Content Data on Prediction Accuracy of Recommender Systems*	Bernd Heinrich Marcus Hopf Daniel Lohninger Alexander Schiller Michael Szubartowicz	Electronic Markets (EM)	This paper is **accepted** and **published** in Volume 31, Issue 2 in the journal *Electronic Markets*.
	Chapter 4.2: *Global Reconstruction of Language Models with Linguistic Rules – Explainable AI for Online Consumer Reviews*	Markus Binder Bernd Heinrich Marcus Hopf Alexander Schiller	European Conference on Information Systems (ECIS)	This paper was **under review** at *30th European Conference on Information Systems* with the former title "Reconstructing the Language Model BERT with Linguistic Rules – Explainable AI for Online Consumer Reviews". After rework, this paper is **submitted** to the special issue on *Explainable and responsible artificial intelligence* of the journal *Electronic Markets* in Mai 2022.

***Table 1-1** Overview of the papers contained in this dissertation.*

2 Data Understanding of Online Consumer Reviews

"All truths are easy to understand once they are discovered.
The point is to discover them."
Galileo Galilei (*1564; †1641)

2.1 Paper: Long-term Sequential and Temporal Dynamics in Online Consumer Ratings

Current Status	**Citation**
This paper is **accepted for publication** in *Proceedings of the 30th European Conference on Information Systems (ECIS 2022)*.	Heinrich, B., T. Hollnberger, M. Hopf and A. Schiller (2022a). "Long-term Sequential and Temporal Dynamics in Online Consumer Ratings" Proceedings of the 30th European Conference on Information Systems (ECIS 2022).

LONG-TERM SEQUENTIAL AND TEMPORAL DYNAMICS IN ONLINE CONSUMER RATINGS

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

Theresa Hollnberger, University of Regensburg, Regensburg, Germany, theresa.hollnberger@ur.de

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

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

Abstract

Online consumer ratings provide important feedback for businesses and yield essential purchase information for consumers. Extant literature has recognized the importance of sequential and temporal dynamics of consumer ratings, but has shed light upon short-term dynamics (e.g., an initial decreasing rating trend) and lacks analyses of long-term dynamics. Existing findings thus cannot explain these long-term dynamics, which are particularly important as many items receive ratings over the long term. In this paper, we therefore examine long-term sequential and temporal dynamics in consumer ratings and in particular whether initial rating dynamics influence average ratings in the long-term. To do so, we apply regression models to an extensive long-term review dataset. First, we find and explain a new long-term sequentially increasing rating trend which leads to a U-shaped relationship between ratings and their order. Second, we reveal that strong initial rating dynamics have significant negative impact on long-term average ratings.

1 Introduction

Vast amounts of online consumer star ratings for items (e.g., products or services) are generated daily on various review platforms and other social media channels. These ratings serve as essential purchase information that influence other potential consumers, give direct consumer feedback for businesses (Helversen et al., 2018) and allow platforms to model consumer preferences for well-personalized item recommendations (Ricci et al., 2015). For instance, a recent consumer survey showed that 87% of consumers read online reviews for local businesses in 2020, that ratings were the most important information of a consumer review, and that only 48% of consumers would consider using a business with an average rating below 4 stars (Murphy, 2020). As Godes and Silva (2012, p. 448) succinctly put it, "information from others matters". A central characteristic of consumer ratings is that they are generated from the item's market launch date on in a timely-ordered manner capturing consumer assessments at different points in time and in (sequential) order. Existing literature has shed light upon sequential and temporal dynamics of online consumer ratings, with dynamics constituted by both (a) trends of consumer ratings depending on their time and order positions and (b) underlying theoretical explanations of these trends (cf., e.g., Li and Hitt, 2008; Li et al., 2019; Park et al., 2021). For instance, Li and Hitt (2008) revealed a highly interesting trend indicating a decrease in the consumer rating depending on the time of the ratings, caused by late adopters disconfirming "too high" ratings of enthusiastic early adopters (self-selection explanation). The analysis of such dynamics is vital to enable a deeper understanding of online consumer ratings.

So far, prior research on sequential and temporal rating dynamics has captured a short-term view on small-size rating datasets (e.g., Godes and Silva (2012) focused only on the first 50 ratings). These works found increasing and decreasing linear rating trends in the short term that could be explained by means of self-selection or motivation of consumers (e.g., Askalidis et al., 2017; Godes and Silva, 2012; Wang et al., 2018). However, most items receive ratings not only in the short term but also over the long term, spanning many years. Indeed, consumers continue to rate items even when hundreds of ratings are already available. Moreover, there is (practical) evidence regarding the importance of long-term rating dynamics: Chevalier and Mayzlin (2006) found significant effects of long-term temporal rating dynamics on book sales by means of a differences-in-differences analysis. Ha et al. (2015) examined the impact of online ratings on book sales in the long term with a hypothesis-driven analysis and reported that "the influence of seller-site

reviews remained significant after a considerable lapse of time" (p. 383). Thus, the understanding of long-term sequential and temporal dynamics is vital, but not facilitated by existing short-term studies. More precisely, it is relevant whether short-term dynamics (e.g., over the first 50 ratings of an item) continue and whether they exhibit any impact on subsequent ratings in the long term. Here, extant literature has already analyzed effects of the very first rating on the trend of the subsequent ratings in the short term (e.g., in the first 20 ratings). While Park et al. (2021) as well as Lederrey and West (2018) find that the first rating influences the trend of the subsequent short-term ratings, Wu and Huberman (2010) state that the average rating converges to that of the dataset regardless of the first rating, which is theoretically explained by disconfirmation. In particular, strong initial rating dynamics (e.g., in the first 20 ratings of an item) often result from "too high" ratings by early adopters (Hu et al., 2017; Li and Hitt, 2008), which also has only been analyzed in the short term. Summing up, extant literature neither analyzes long-term rating dynamics (e.g., whether a rating trend reversal occurs over time) nor the impact of initial rating dynamics on the long-term average rating. In this paper, we therefore examine the following two research questions:

1. What are the long-term sequential and temporal dynamics in online consumer ratings?
2. What is the impact of initial rating dynamics on the long-term average rating of items?

We examine these research questions by means of regression models and statistical tests analyzing an extensive review dataset, spanning 13 years with more than 1.9 million restaurant reviews in total and a median above 350 ratings per item. Our results support the existence of decreasing trends in consumer ratings depending on both time and order, caused by disconfirmation, which is in line with extant literature (e.g., Bjering et al., 2015; Li and Hitt, 2008) – but only in the short term. In the long term, we find a new sequential rating trend that suggests the existence of a *second disconfirmation*, resulting in an *in*creasing rating trend for high order positions. Similarly, our analysis of initial rating dynamics reveals essential novel findings for the long term. That is, strong initial rating dynamics (e.g., in the first 20 ratings of an item) have a significant negative impact on the long-term average rating of items. In particular, items initially receiving high ratings exhibit the lowest long-term average rating, indicating a *lasting (negative) effect based on initially stimulated expectations*.

The remainder of the paper is structured as follows. In the next section, we position our work in literature and discuss the research gap. Thereafter, we introduce the dataset used to analyze the long-term dynamics in online ratings. Subsequently, our analysis based on regression models and statistical tests follows. In the penultimate section, we discuss our findings in theory and point out implications for practice. Finally, we summarize, reflect upon limitations and present an outlook on future research.

2 Background and Related Work

In this section, we position our research in the field of rating dynamics in online consumer reviews. We discuss extant research which analyzes the sequential and temporal rating dynamics in general and in particular, the initial ratings and their impact on subsequent ratings. For this purpose, we first searched the databases ACM Digital Library, GoogleScholar, IEEE Xplore, ScienceDirect, SpringerLink and Wiley, following a standard approach to prepare the related work (Levy and Ellis, 2006). We used the following two search queries: I. *(sequential OR temporal) AND dynamics AND "online reviews" AND rating* and II. *("first rating" OR "initial rating" OR "early adopter") AND "online reviews" AND ("long-term rating" OR "mean rating" OR "average rating")*. Since the number of results in some databases exceeded the manageable scope, we restricted the search to the first 150 results for each search term and database. In sum, the search led to 519 publications, which were manually screened based on title, abstract and keywords, and if necessary, a more detailed analysis. All papers that used a dependent variable other than (average) rating, that relied solely on autoregression analysis or that used sequential and temporal variables only as control variables are not within our scope and thus were excluded from examination. This way, we identified 9 relevant papers, based on which we conducted an additional forward and backward search. After all, 11 papers were identified as highly relevant for our work.

We begin our review with works studying sequential and temporal rating dynamics in general and provide a discussion of the different explanations that have been proposed for the analyzed rating trends. A first theoretical motivation is the *diagnosticity assessment explanation*, which is suggested by the seminal work of Godes and Silva (2012) to explain a sequentially decreasing trend in ratings. According to this explanation, consumers are less able to determine which reviews are relevant for them with an increasing number of reviews, which may lead to an information overload problem. Since only the first 50 rating

positions in the rating sequence (order) are focused on with a linear ordered-logit model in this paper, the decreasing trend as well as the corresponding diagnosticity assessment explanation are only discussed for a short-term view. Guo and Zhou (2016) use the same explanation to motivate the relationship between the previous average rating and the subsequent rating, while considering items with a mean of 718 ratings. Yet, they only analyze the moderating effect of order (volume of prior reviews) on this relationship, thus disregarding long-term sequential dynamics. Wang et al. (2018) also theoretically ground their work on the diagnosticity assessment explanation. As additional theoretical motivation, they also consider the effects of rating heterogeneity as well as the diagnostic abilities of the consumer on ratings. According to this, high diagnostic abilities and low heterogeneity are associated with an increasing trend, while low diagnostic abilities and high heterogeneity are associated with a decreasing trend. Thus, based on this extended explanation, the authors argue that both in- and decreasing trends can be analyzed in the short term. However, these results are based only on a very small dataset (2,595 reviews in total), limiting their validity. Another explanation for the sequentially decreasing rating trend in the short term is motivated by the dissimilarity of reviewers (Godes and Silva, 2012). This explanation states that the similarity between reviewers decreases over the review sequence, which can also negatively affect the rating assessments (Godes and Silva, 2012). It is closely related to another explanation, the *self-selection explanation,* which expresses that consumers self-select when entering the market for an item (Bjering et al., 2015; Li and Hitt, 2008). Thus, dissimilar consumers (or consumer groups) reflect their individual expectations over time resulting in disconfirming ratings (discussed in more detail below). In contrast to these works, Li et al. (2019) identify an increasing sequential rating trend, with the theoretical motivation being consumers tending to leave higher ratings for popular items (measured by the number of reviews). However, analogous to the other works mentioned so far, only a linear model is applied, which is unable to detect whether a trend reversal occurs. In addition, their description of the dataset does not clarify how many reviews per item were analyzed and therefore the robustness and the validity of their results remain unclear. Furthermore, the *motivation-based explanation* of Wu and Huberman (2008) is often cited as a reason for the decreasing trend over time and order in the short term (Askalidis et al., 2017; Godes and Silva, 2012). This explanation states that the motivation to write a review is greater if the own assessment differs strongly from prior ratings, especially if these ratings are perceived as "too high". This is because the perceived value of the review then outweighs the costs of writing it (Wu and Huberman, 2010, 2008). Finally and with a focus on temporal dynamics, the *macro explanation* (Godes and Silva, 2012) is grounded on consumers generally becoming more critical over time. Therefore, a decreasing trend over time results in the short term. Summing up, extant work has found interesting (linear) sequential and temporal dynamics of online consumer ratings but only in a short-term view.

With regard to the second research question, we discuss literature dealing in particular with the initial rating dynamics (e.g., in the first 20 ratings of an item) and their impact on subsequent ratings. Thereby, the first ratings or the very first rating are referred to as the cause of increasing or decreasing trends. As introduced above, a decreasing trend in the short term can be explained by *self-selection*, as different consumer groups review an item at different points in time (Li and Hitt, 2008), resulting in disconfirming ratings. This explanation is supported by the Expectancy-Disconfirmation Theory (Oliver, 2014), stating that failed expectations entail negative disconfirmation. The theoretical motivation behind this is that the first ratings are usually left by enthusiastic early adopters who, for example, are very eager to experiment or already have a positive attitude towards the item. This means that these ratings of an item are usually "too high" and biased (for certain items the opposite can also occur, i.e., the ratings are "too low"). Since early adopters' ratings generally do not correspond to the assessment of most late adopters, the latter consumers tend to react with disconfirmation (Li and Hitt, 2008). Thereby, they give even lower ratings and undershoot the rating they would have given without the bias of these too high ratings, which leads to a strongly decreasing rating trend in the short term (Li and Hitt, 2008). Thus, this decreasing trend is explained by self-selection (Bjering et al., 2015). These works, however, focus on the impact of the initial ratings in the short term and thus, do not consider long-term dynamics. A specialization of these analyses is provided by Park et al. (2021), who focus on the impact of the first rating on subsequent ratings. Their work is theoretically motivated by the information-availability bias, stating that when a product receives a negative first review, it suffers low initial sales and cannot not fully overcome the negative bias created by the first rating. Indeed, the average rating twelve months after a negative first rating still is lower than after a positive first rating for the same product on different platforms (Park et al., 2021). A similar analysis is conducted by Lederrey and West (2018), who theoretically ground their work on the thesis that customers may be biased to follow other customers' previous ratings of the same product. Their research reveals that the first

rating has an impact until the item has received 20 ratings. Contrary findings are reported by Wu and Huberman (2010), who state that the average rating converges to that of the dataset regardless of the first rating, which is theoretically motivated by disconfirmation. Summing up, extant works focus on the first ratings and its impact on the subsequent ratings in the short term, but do not analyze the impact of initial rating dynamics on long-term ratings.

Overall, existing literature has revealed interesting trends theoretically motivated by different explanations. A particularly notable and frequently referenced theoretical motivation is given by the self-selection explanation, which is grounded on Expectation-Disconfirmation Theory. Existing works thus provide valuable contributions analyzing sequential and temporal rating dynamics. However, none of these works focus on long-term dynamics (1). Moreover, all works apply only linear models and thus are unable to analyze whether a trend reversal occurs at any time or order position (2). Previous works discussing the impact of initial ratings on the average rating also restrict their analysis to impacts on the short term (3). In addition, these works do not focus on initial rating *dynamics* (4). In this work, we aim at addressing the identified research gap by analyzing and theoretically motivating long-term (ad 1) sequential and temporal dynamics striving to also examine non-linear dynamics (ad 2) as well as the impact of initial rating dynamics (ad 4) on the long-term average rating (ad 3).

3 Data

In this section, the dataset used to analyze the long-term dynamics in online consumer ratings is described in detail. The dataset originates from an established online platform for local business reviews. In particular, the dataset consists of 2,396,643 reviews for items (restaurants, bars, and cafés) in New York City, which were written in the time period from 2004 to 2016. For each review, the date of creation, the item and user ID as well as the star rating (given on a five-tier scale from 1 star to 5 stars) is stored. It can be assumed that our dataset does not include fake reviews, as all reviews of the dataset were examined by the providing online platform and explicitly labeled as trustworthy.

For the analysis of the sequential and temporal dynamics of ratings R_{ui}, the following four variables are determined for a review rev_{ui} of the user u for the item i: $TIME_{ui}$, $ORDER_{ui}$, $REVAVG\_U_{ui}$ and $REVAVG\_I_{ui}$ (cf. also Godes and Silva, 2012). The first variable $TIME_{ui}$ denotes the number of days between the first review of the item i and the review rev_{ui}. Thus, this variable allows to analyze temporal dynamics. The second variable $ORDER_{ui}$ describes the position of a review rev_{ui} in the review-sequence (ascendingly ordered by date of creation) of an item i and can be used to examine sequential dynamics. Since in the dataset at hand, the date of creation is only given by the day but not by the exact time, it is not possible to give different sequence positions to the reviews of an item that were written on the same day. For this reason, reviews of an item with the same date of creation receive the same order position in the review-sequence and thus, the same value for $ORDER_{ui}$. The variable $REVAVG\_U_{ui}$ is used to account for the average rating of the user u, as some users are more positive or negative than others. For this, the value of the variable is calculated as the average rating over all reviews that a user u has written, excluding the review rev_{ui}. Reviews of users who have only written one review are removed from the dataset, because it is not possible to determine a value for the variable $REVAVG\_U_{ui}$. In this way, 305,420 reviews are removed from the dataset (however, to not interfere with sequential dynamics, these reviews are nevertheless considered for determining the values of $ORDER$). Finally, $REVAVG\_I_{ui}$ is used to account for the average rating of the item i and is calculated analogously to $REVAVG\_U_{ui}$.

					Spearman's Rank Correlations with			
	Mean	Std. Dev.	Min.	Max.	TIME	ORDER	REVAVG_U	REVAVG_I
Rating	3.76	1.17	1	5	-0.02	0.06	0.24	0.32
TIME	1,646.35	1,084.86	1	4,374		0.58	0.03	-0.08
ORDER	354.99	630.44	1	8,222			0.06	0.18
REVAVG_U	3.75	0.69	1.00	5.00				0.06
REVAVG_I	3.74	0.43	1.30	4.97				

Table 1. Dataset summary statistics.

That is, the value of $REVAVG\_I_{ui}$ equals the average rating of the item i without the review rev_{ui} of user u. Since a sufficiently large number of reviews per item is necessary for a well-founded analysis of long-term dynamics, only items that have received at least 50 reviews are considered. This leads to the

removal of another 165,001 reviews from the dataset. Finally, the adjusted dataset consists of 1,926,222 reviews from 278,389 users who wrote at least two reviews and that belong to 8,509 items that received at least 50 reviews. Table 1 shows the summary statistics of the four determined variables and the rating as well as the Spearman's Rank Correlations between the variables.

To get a first model-free view on the sequential and temporal trends of the online ratings, the plots in Figure 1 illustrate the relationships between the (cumulative) average rating and the variables $TIME$ or $ORDER$. Here, a point represents the (cumulative) average rating of all reviews at the same $TIME$ or $ORDER$ position. To support robust results, we restrict our analysis to $TIME \leq 3{,}000$ and $ORDER \leq 700$, as within these limits the number of reviews used to calculate the average rating for one point is always above 200. Thus, the reviews with $TIME \leq 3{,}000$ and $ORDER \leq 700$ represent our long-term view. The reviews with $TIME \leq 365$ and $ORDER \leq 50$ represent our short-term view (in line with the definitions of Godes and Silva, 2012; gray area in Figure 1).

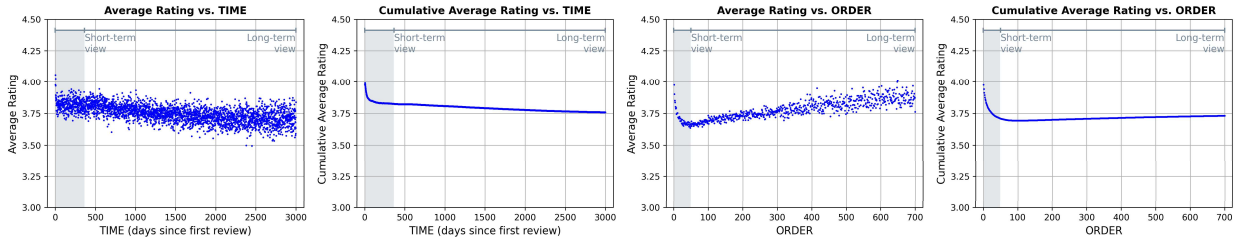

Figure 1. Model-free view on the sequential and temporal rating trends in our dataset.

In the short-term view, the plots show a strong initial decreasing trend of the (cumulative) average rating for increasing $TIME$ and $ORDER$ positions. The long-term view illustrates that the (cumulative) average rating continues to decrease for $TIME$ positions > 365. In contrast, the trend for $ORDER$ is different. Here, the average rating starts to increase for $ORDER$ positions >~ 50 and the cumulative average rating starts to increase for $ORDER$ positions >~ 100.

4 Long Term Dynamics in Online Ratings

We begin this section by analyzing short-term sequential and temporal dynamics to establish a link to extant works and then, we carefully extend this analysis to the long-term view. After that, a detailed analysis is presented to determine the impact of initial rating dynamics on the long-term average rating. The section concludes with a examination for robustness of the results.

4.1 Long-term sequential and temporal dynamics in online ratings

We analyze the sequential and temporal dynamics in online ratings by means of ordered-logit regression models (McCullagh, 1980). In particular, we chose this type of regression model because the dependent variable (the rating) is discrete and ordered. More precisely, we model the rating $R_{ui} \in \{1,2,3,4,5\}$ assigned by user u for item i as a function of the independent variables $TIME_{ui}$, $ORDER_{ui}$, $REVAVG\_U_{ui}$ and $REVAVG\_I_{ui}$ described in Section 3. Thus, the linear ordered-logit model (based on Godes and Silva, 2012), estimated using maximum likelihood, is as follows:

$$R_{ui}^{*} = \beta_1 * TIME_{ui} + \beta_2 * ORDER_{ui} + \beta_3 * REVAVG\_I_{ui} + \beta_4 * REVAVG\_U_{ui} + \varepsilon_{ui'} \quad (1)$$

$$\begin{aligned} R_{ui} = 1 &\Leftrightarrow R_{ui}^{*} < \mu_1, \\ R_{ui} = k \in \{2,3,4\} &\Leftrightarrow R_{ui}^{*} \in [\mu_{k-1}, \mu_k) \; and \\ R_{ui} = 5 &\Leftrightarrow R_{ui}^{*} \geq \mu_4. \end{aligned} \quad (2)$$

Here, both the coefficients $\beta_1, \ldots, \beta_4$ of the independent variables and the thresholds $\mu_1, \ldots, \mu_4$ are estimated. With the latter, it is possible to transform the latent evaluation of the underlying linear model $R_{ui}^{*} \in \mathbb{R}$ to the discrete ordinal response variable R_{ui}. For example, $R_{ui} = 1$ results if the determined value R_{ui}^{*} is smaller than the threshold μ_1. With $\varepsilon_{ui'} \sim N(0,1)$, the random error term is considered.

We aim to analyze whether different sequential and temporal dynamics occur over the long term compared to the short term. Thus, the model is applied once for the whole long-term dataset (1,508,947 reviews) yielding the long-term model (LTM) as well as for the restricted short-term dataset (166,011 reviews) yielding the short-term model (STM). The results of the STM and the LTM are presented in Table 2. The estimated coefficients for $TIME$ and $ORDER$ are negative for both models, with their size varying

depending on short-term vs. long-term view. Considering the absolute values, the coefficient for $TIME$ for the STM (-0.0115) is lower than the one for the LTM (-0.0191). For $ORDER$, however, it is much higher (-0.1301 vs. -0.0122). Furthermore, the coefficients are almost all significant at the 0.001 level. Only the coefficient for $TIME$ of the STM is significant at the 0.05 level.

The results show that the short-term sequential and temporal dynamics discovered by many previous works (cf. Section 2) also hold true for our dataset. In particular, the average rating decreases in the short-term view. However, the results also support our suggestion that long-term dynamics may differ considerably from short-term dynamics. Indeed, the coefficient for $ORDER$ differs by a magnitude of more than 10 between STM and LTM, and the coefficient for $TIME$ also varies. In line with the discussion in Section 3, this may indicate that a steady linear decrease of the rating with respect to $TIME$ and $ORDER$ does not hold true over the long term. Moreover, the linear ordered-logit model discussed in literature so far seems to be inappropriate to carefully model the dynamics over the long term.

	STM			**LTM**			**QLTM**		
	Coef.	S.E.	95% CI	Coef.	S.E.	95% CI	Coef.	S.E.	95% CI
TIME	-0.0115*	0.005	[-0.02, -0.00]	-0.0191***	0.002	[-0.02, -0.02]	-0.0190***	0.002	[-0.02, -0.02]
TIME²							0.0185***	0.002	[0.02, 0.02]
ORDER	-0.1301***	0.005	[-0.14, -0.12]	-0.0122***	0.002	[-0.02, -0.01]	-0.0379***	0.003	[-0.04, -0.03]
ORDER²							0.0191***	0.001	[0.02, 0.02]
REVAVG_I	0.7053***	0.005	[0.70, 0.71]	0.6644***	0.002	[0.66, 0.67]	0.6656***	0.002	[0.66, 0.67]
REVAVG_U	0.4312***	0.005	[0.42, 0.44]	0.4470***	0.002	[0.44, 0.45]	0.4471***	0.002	[0.44, 0.45]
N	166,011			1,508,947			1,508,947		
AIC	437,167			4,053,007			4,052,623		
BIC	437,247			4,053,105			4,052,745		
Nagelkerke Pseudo-R²	0.1794			0.1706			0.1708		

*Table 2. Model coefficients (Coef.) with standard errors (S.E.) and 95% confidence intervals (CI) rounded to two decimals. The asterisks indicate the significance level of the coefficients with the notation: '***': $p \in [0, 0.001]$, '*': $p \in (0.01, 0.05]$.*

Thus, to address these issues, the linear model is extended by polynomial terms of degree two for $TIME$ and $ORDER$ (cf. Equation (3)) allowing to examine whether there exist other, non-linear trends in the long-term dynamics. In this way, a quadratic long-term model (QLTM) is formed and applied to the long-term dataset. The transformation to ordinal ratings is performed as above by Equation (2).

$$R^*_{ui} = \beta_1 * TIME_{ui} + \beta_2 * TIME^2_{ui} + \beta_3 * ORDER_{ui} + \beta_4 * ORDER^2_{ui} + \beta_5 * REVAVG\_I_{ui} + \beta_6 * REVAVG\_U_{ui} + \varepsilon_{ui'} \quad (3)$$

As a polynomial regression (Dean et al., 2017; Rawlings et al., 1998), the QLTM enables to model a non-linear relationship between rating and $TIME$ or $ORDER$. The results in Table 2 show that all coefficients in the QLTM are significant at the 0.001 level. This provides a first indication that the QLTM may be more suitable to model the relationship between rating and $TIME$ or $ORDER$ than the linear model proposed by extant literature. In addition to the model coefficients, three measures of quality are determined for each model: Akaike Information Criterion (AIC) (Akaike, 1973), Bayesian Information Criterion (BIC) (Schwarz, 1978) and Nagelkerke Pseudo-R^2 (Nagelkerke, 1991). The first two measures are used to evaluate different models on the same dataset (Kuha, 2004). Both include a penalty term for the number of parameters (Akaike, 1973; Schwarz, 1978), such that the model with lower AIC/BIC has a higher probability of being closer to the true model (Burnham and Anderson, 2004; Raftery, 1995). The AIC and BIC values of the QLTM compared to those of the LTM decrease, whereas the Nagelkerke Pseudo-R^2 increases (cf. Table 2). This means that, on the long-term dataset, the QLTM has a higher probability of being closer to the true model than the linear model. Please note that this holds true even though the differences in AIC and BIC may appear small in relative terms, as the absolute difference is relevant and indicates strong evidence (Burnham and Anderson, 2004; Raftery, 1995).

On this basis, we aim for a statistical analysis of the long-term rating dynamics, focusing on $ORDER$ at first. Since the average rating with respect to $ORDER$ increases after it has initially decreased (cf. Figure 1 and the positive coefficient of $ORDER^2$ in the QLTM), the long-term rating dynamics for $ORDER$ might be described by a U-shaped relationship. A U-shaped relationship means that the effect of an independent variable on a dependent variable is negative for low values, but positive for high values, or vice versa. In our case, a substantiated U-shaped relationship would mean that the sequential rating dynamics change depending on the size of $ORDER$, with $ORDER$ having a negative effect on the rating at first, but a positive effect later on. To test a U-shaped relationship, quadratic regression analysis with a significant quadratic coefficient is not sufficient, because this can lead to false interpretations in case of non-quadratic functions. Instead, the "two-lines test" by Simonsohn (2018) is suitable to test whether a relationship is described by a U-shape. For this purpose, a breakpoint is determined according to Muggeo (2003) and Muggeo (2008) and an interrupted regression model is estimated according to Simonsohn (2018). The corresponding regression model for the test of the relationship between $ORDER$ and the rating R_{ui} is as follows:

$$R_{ui} = \begin{cases} \beta_0 + \beta_1 * ORDER_{low,ui} & if\ ORDER_{ui} < ORDER_c \\ \beta_0 + \beta_2 * ORDER_{high,ui} & if\ ORDER_{ui} \geq ORDER_c \end{cases} \quad (4)$$

Here, the variables $ORDER_{low}$ and $ORDER_{hig}$ are calculated as the difference between $ORDER$ and the breakpoint $ORDER_c$. The results of this test show that the coefficient of $ORDER_{low}$ is negative (-0.0134) and significant at the 0.001 level and the coefficient of $ORDER_{high}$ is positive (0.0004) and significant at the 0.001 level, with the breakpoint $ORDER_c$ at 17. Thus, the two-lines test according to Simonsohn (2018) supports the supposed U-shaped relationship between rating and $ORDER$. Although the coefficient of $ORDER_{high}$ may appear to be small, it actually has a distinctive effect on the ratings, as the value of $ORDER_{high}$ can be very large (e.g., 600) compared to ratings (ranging from 1 to 5).

Moreover, we conducted a similar analysis for $TIME$. More precisely, we applied the two-lines test of Simonsohn (2018) for $TIME$ in the same way as for $ORDER$. This yields that the breakpoint $TIME_c$ is 12 and that both coefficients, the one for $TIME_{low}$ and the one for $TIME_{high}$, are negative (-0.0150 and -0.0001) and significant at the 0.001 level. Therefore, the test refutes that the relationship between rating and $TIME$ may be described by a U-shape. Rather, the rating continues to decrease after the breakpoint, but more slightly than before, since the absolute value of the coefficient for $TIME_{high}$ is lower than that for $TIME_{low}$.

4.2 Impact of initial rating dynamics on the long-term average rating

In this section, we examine whether there exists an impact of initial rating dynamics on the long-term average rating of items. Considering the ratings of all 8,509 items (as illustrated in Figure 1), the initial rating dynamics in our dataset exhibit a decreasing trend, in line with extant works (cf. Section 2). We first analyze whether this trend also holds for each individual item by analyzing the slope of the first 20 ratings for each item. We used the slope of the initial ratings instead of the average rating, as the slope allows to quantitatively assess rating dynamics. The choice of 20 was motivated by the breakpoint of the two-lines test (cf. Section 4.1), but results for other choices are similar (cf. Section 4.3). The slopes are in the range [-0.18, 0.16] with mean -0.01 and standard deviation 0.04, which reveals that there are not only items with decreasing initial trends but also with increasing or constant initial trends.

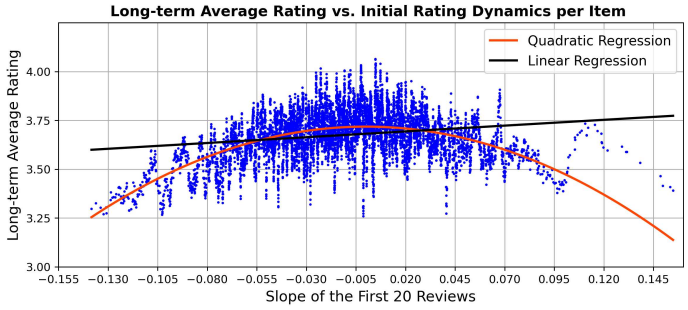

Figure 2. Illustration of the relationship between initial rating dynamics and long-term average rating.

Based on this, the impact of the initial rating dynamics on the long-term average rating LTR_i of an item i

(i.e., the average of all ratings R_{ui} of i in the long-term view) will be examined. As discussed above, one could assume that the characteristic of the initial rating dynamics (represented by the initial slope $SLOPE_i$) for an item i has a very limited impact on its long-term average rating LTR_i, since these dynamics concern only the initial 20 ratings. However, the plot in Figure 2 indicates that there indeed exists an interesting non-trivial relationship. Each point in the plot represents the moving average of the long-term average rating within a window size of twenty (the moving average was used for illustration purposes only, the following regression models are applied to the actual observations per item).

Based on the plot, we examine this relationship by means of a linear regression model (5) and a non-linear, quadratic regression model (6), which are also illustrated in Figure 2.

$$LTR_i = \beta_0 + \beta_1 * SLOPE_i \quad (5)$$

$$LTR_i = \beta_0 + \beta_1 * SLOPE_i + \beta_2 * SLOPE_i^2 \quad (6)$$

Here, β_0, β_1 and β_2 are the estimated coefficients. The results of the models are summarized in Table 3. They suggest that the quadratic regression model is much better suited to describe the relationship between $SLOPE_i$ and LTR_i. Indeed, the values of AIC and BIC support the superiority of the quadratic regression model with very strong evidence (Burnham and Anderson, 2004; Raftery, 1995). Similarly, the value for the Adjusted R^2 (Wherry, 1931; Yin and Fan, 2001) is much higher for the quadratic regression model. The linear coefficient for $SLOPE$ of the quadratic regression model is not significant and thus the vertex of the resulting parabola is not significantly different from zero. In addition, the coefficient for $SLOPE^2$ is negative and significant with a p-value below 0.001, which indicates that the parabola is inverted. Therefore, following this regression model, items with an initial slope close to zero reach the highest long-term average rating, and as the slope increases in absolute terms the items reach a lower long-term average rating. In other words, items with a strong increasing or decreasing initial trend achieve a lower long-term average rating than items that do not exhibit such a trend.

	Linear Regression Model	**Quadratic Regression Model**
Intercept	3.6815***	3.7179***
SLOPE	0.5934***	0.0031
SLOPE²		-24.2109***
N	8,509	8,509
AIC	11,074	10,927
BIC	11,096	10,955
Adjusted R²	0.0026	0.0198

*Table 3. Model results. '***' indicates $p \in [0, 0.001]$.*

We further verify whether the relationship between initial slope and long-term average rating has a significant inverted U-shape by again applying the two-lines test (Simonsohn, 2018) as described in Section 4.1 (see Equation (4)). This yields the breakpoint $SLOPE_c$ at -0.0042, the positive coefficient 2.7260 for $SLOPE_{low}$ and the negative coefficient -2.2655 for $SLOPE_{high}$, both significant at the 0.001 level. Thus, the test supports a significant inverted U-shaped relationship.

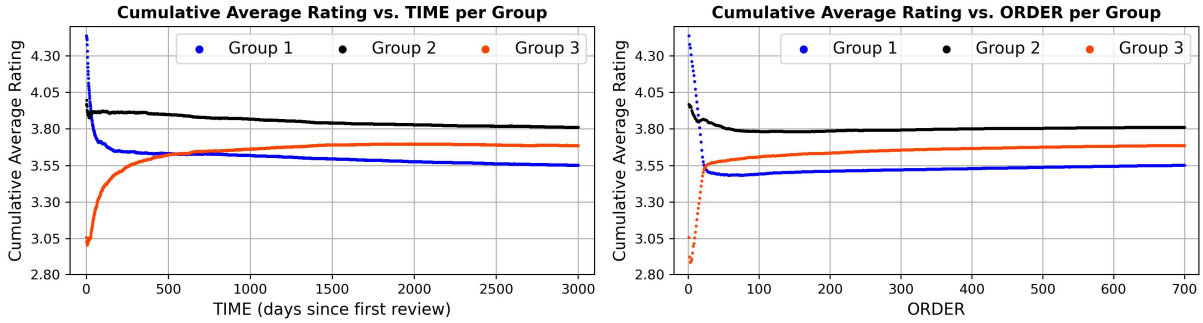

Figure 3. Illustration of cumulative average rating vs. TIME and ORDER per group.

To illustrate these results further, we define three groups of items. Group 1 consists of the 500 items with the most negative slopes, Group 2 comprises the 500 items with the lowest slopes in absolute terms and

Group 3 contains the 500 items with the most positive slopes. For every group, the change in their cumulative average rating over $TIME$ and $ORDER$ is shown in Figure 3.

The cumulative average rating in Group 1 and Group 3 changes in the opposite direction, since it is initially high in Group 1 and then decreases over $TIME$ and $ORDER$ and it is initially low in Group 3 and then increases. Here, it is especially remarkable that the long-term average rating (3.6864) of Group 3 is higher than the one of Group 1 (3.5514). However, the highest long-term average rating is achieved by Group 2 (3.8110).

4.3 Robustness

Both the long-term sequential and temporal rating dynamics presented in Section 4.1 and the impact of initial rating dynamics on the long-term average rating in Section 4.2 have been assessed in different ways. Indeed, the long-term sequential and temporal dynamics were firstly explored by a model-free view of the data (cf. Figure 1) and then analyzed by means of a quadratic regression model exhibiting statistically significant quadratic terms. As a first indication of robustness of this result, strongly improved values for AIC/BIC/Nagelkerke Pseudo-R^2 of the quadratic regression model compared to the linear regression model were revealed (cf. Table 2). Moreover, the results, especially the U-shaped relationship, were statistically supported by the two-lines test. Similarly, the impact of initial rating dynamics on the long-term average rating have been illustrated (cf. Figure 2 and Figure 3), analyzed by means of a regression model and related measures (cf. Table 3) and supported by the two-lines test. Still, we assess the robustness of our results in further specifications. In particular, we discuss both (a) data-related robustness and (b) model-related robustness.

Ad (a): In our analysis we restricted our dataset in multiple ways, which led to the exclusion of 887,696 reviews from our initial dataset (cf. Section 3). To analyze whether this exclusion affects our results, we take into account all available reviews, leading to a dataset with 2,236,368 reviews of 9,272 items. Only the restriction to items with at least 50 reviews, which is necessary for an analysis of long-term dynamics, remains. On this dataset, we firstly constructed a linear and a quadratic regression model in line with Section 4.1. The results were analogous: Again, all coefficients in the quadratic model are significant at the 0.001 level, have the same sign as in Section 4.1, and AIC, BIC and Nagelkerke Pseudo-R^2 are strongly in favor of the quadratic model, supporting our results from Section 4.1. Second, we also assessed the robustness of the results in Section 4.2 on this extended dataset. Here, the slope of the first 20 ratings was in the range [-0.23, 0.17], with a mean of -0.01 and a standard deviation of 0.05. Again, the coefficient for the linear term ($SLOPE$) in the quadratic regression model is not significant, while the coefficient for the quadratic term ($SLOPE^2$) is negative and significant at the 0.001 level, AIC/BIC/Adjusted R^2 strongly speak in favor of the quadratic regression model, and the significant inverted U-shape of the relationship between rating and slope is supported by the two-lines test. This shows the robustness of our results in Section 4.2 regarding dataset restrictions.

An alternative possible explanation for the results presented in Section 4.1 would be that items that receive a larger number of ratings overall tend to receive better ratings, as suggested by Li et al. (2019). To verify this explanation, we repeated the analysis using three particular datasets, whereby we restricted these datasets such that they contained only reviews of items with less than 200, 400 or 600 reviews, respectively. The results were analogous to the ones presented in Section 4.1: Again, all coefficients in the QLTMs are significant at the 0.001 level, have the same sign as in Section 4.1, AIC, BIC and Nagelkerke Pseudo-R^2 are strongly in favor of the quadratic models, and the two-lines test yields a significant U-shape in each case. Similarly, we assessed the robustness of the results in Section 4.2 in this regard. The results show that the value of the intercept in the quadratic regression model indeed increases slightly with an increasing number of reviews (3.6883 for less than 200 reviews, 3.6970 for less than 400 reviews and 3.7024 for less than 600 reviews), supporting that a larger number of ratings corresponds to better ratings. Yet, the results presented in Section 4.2 still hold true in each case. Indeed, the coefficient for $SLOPE^2$ is negative and significant at the level 0.001, AIC/BIC/Adjusted R^2 speak in favor of the quadratic regression model, and the significant inverted U-shape of the relationship is supported by the two-lines test in each case. Thus, this analysis also supports the robustness of the results in Section 4.2 and shows that our result is not conflicting, but complementary to the explanation of Li et al. (2019). To further ensure that the results in Section 4.1 are independent of the choice of thresholds for $TIME$ and $ORDER$ determining the short-term view, other thresholds partly based on extant literature (Li and Hitt, 2008) were evaluated. In particular, we used four additional thresholds for $TIME/ORDER$ (126/62, 365/100, 365/150, 730/150) to set the

short-term view. This analysis showed that all coefficients had the same sign and are significant at least at the 0.01 level regardless of the used thresholds. Thus, these results were in line with the those results presented in Section 4.1, supporting their robustness.

Additionally, we also assessed the robustness of the group-based analysis in Section 4.2 (even though this analysis is for illustration purposes). Here, we repeated the analysis using 1,000 items in each group (instead of 500). In line with Section 4.2, the cumulative average rating of Groups 1 and 3 changes in the opposite direction, and the highest long-term average rating is achieved by Group 2 (3.82) followed by Group 3 (3.69) and Group 1 (3.61). Overall, this ensures the data-related robustness of our results.

Ad (b): To analyze the model-related robustness of our results in Section 4.1, we verified our results by means of another model and additional model terms. Instead of the ordered-logit model, an ordered-probit model is also in general suitable for our data (cf. Binder et al., 2019; McKelvey and Zavoina, 1975). The signs and the significance of all coefficients of the ordered-probit model for STM, LTM and QLTM were consistent to the results presented in Section 4.1. Furthermore, we also extended the presented QLTM by additional polynomial terms. In particular, we constructed models with terms up to degree 4 for both $TIME$ and $ORDER$. Quadratic terms for both were significant in each of the models, at least at the 0.05 level. In particular, neither the non-linear terms of $TIME$ nor of $ORDER$ could be replaced by higher polynomials of the other variable. These results further support that there indeed is a long-term non-linear relationship between the ratings and both $TIME$ and $ORDER$.

Finally, we analyzed whether the choice of 20 for the number of ratings used to calculate the slope of the initial ratings had an effect on the results in Section 4.2. To this end, we conducted analogous analyses using the first 10, 17 (i.e., the breakpoint obtained in Section 4.1) and 30 ratings for calculating the slope. The results are completely in line with the results presented in Section 4.2 in each case: Again, the coefficient for the linear term ($SLOPE$) in the quadratic regression models is not significant, while the coefficient for the quadratic term ($SLOPE^2$) is negative and significant at the 0.001 level, AIC/BIC/Adjusted R^2 speak in favor of the quadratic regression models, and the two-lines tests yield a significant inverted U-shaped relationship. This shows the model-related robustness.

5 Discussion and Implications

Our analysis yields two novel key findings for long-term dynamics in online consumer ratings, which extend existing knowledge, thus allowing a deeper understanding of online consumer ratings. In this section, we discuss these findings in theory and outline implications for practice.

First, the long-term relationship between the sequential order of reviews and their ratings is U-shaped with an increasing rating trend for ratings of high order positions, indicating a second disconfirmation.

Indeed, in line with literature, our results support an overall *decreasing sequential and temporal trend* of ratings in the *short term*. Reconciling extant literature, this trend can be explained as follows: Due to self-selection and dissimilarity of reviewers, enthusiastic early adopters leave high ratings at the beginning followed by a subsequent disconfirmation of further reviewers leaving lower ratings (in line with Bjering et al., 2015; Li and Hitt, 2008). In particular, these late adopters tend to give even lower ratings and undershoot the rating they would have given without the bias of these too high ratings, which leads to a decreasing rating trend in the short term (cf. Ho et al., 2017; Li and Hitt, 2008). This decreasing trend is exacerbated by a higher motivation to write a review with a strongly different opinion compared to existing ratings (in line with Wu and Huberman, 2008). Beyond that and as significant extension of the existing body of knowledge, we are the first to show that this short-term decreasing rating trend changes to an *increasing sequential trend* in the *long term*, resulting in a U-shaped relationship of the sequential order and consumer ratings. In particular, this is shown by the QLTM as well as the two-lines test (cf. Section 4.1). This relationship can be explained by a *second disconfirmation* in the long term, which has not been recognized by extant literature. More precisely, the second disconfirmation is similar to the first, but in this case, it is not the early adopters with too high ratings causing the effect, but the subsequent (overcritical) reviewers that undershot ratings. Hence, the reviewers in the long term assess these subsequent ratings from the short term as too low and disagree with the undershooting reviewers, leading to an increasing long-term sequential rating trend. Indeed, this explanation is sustained by the Expectancy-Disconfirmation Theory (Oliver, 2014), which states that when the assessed item performance differs from the consumer's expectations, both negative disconfirmation (if performance fails to meet expectations) and positive disconfirmation (if performance exceeds expectations) may result. Notably, as extant literature has

focused on a short-term view, this second disconfirmation is the first explanation for a sequential trend in the long term and in particular extends the short-term disconfirmation discussed in literature to the long term. Further, it is interesting that time still exhibits a decreasing rating trend also in the long term. This might seem peculiar at first glance, but is not necessarily surprising as time and order are indeed partly related but not (perfectly) dependent variables (cf. the discussion in Godes and Silva, 2012). In particular, these different trends are possible, since ratings with high order positions can occur at low time values and vice versa. The explanation for a decreasing rating trend over time is that consumers are more critical with older items (cf., e.g., Godes and Silva, 2012). Our work indicates this dynamic to also hold true in the long term, even though it gets weaker (cf. two-lines test for time in Section 4.1).

Second, initial rating dynamics have a significant impact on the long-term average rating of items, particularly indicating that "initially controversial" items achieve a lower long-term average rating, reasoned by a lasting effect based on initially stimulated expectations.

Indeed, our quadratic regression model as well as the two-lines test (cf. Section 4.2) show that strong initial rating dynamics have a significant negative impact on the long-term average rating of items, which extends the present state of knowledge from extant literature (Lederrey and West, 2018; Li and Hitt, 2008; Park et al., 2021). More precisely, we find that items with strong initial in- or decreasing rating trends tend to obtain a lower long-term average rating compared to other items. This indicates that items which receive strong initial disconfirmation – indicated by in- or decreasing rating trends – are also assessed in a more critical manner in the long term, resulting in lower long-term average ratings. Here, it is particularly interesting that items with very high first ratings, but a strongly decreasing initial rating trend tend to exhibit the lowest long-term average ratings. Thus, the consumer assessments of such a controversial item do not completely recover from the strong initial disconfirmation represented by the negative initial rating trend. This can be further explained as follows: Controversial items (especially those with zealous early adopters) exhibit excessively positive, enthusiastic reviews, raising the expectation of future consumers to unrealistically high standards even in the long term. Thus, there is a large number of consumers for whom the actual item performance fails to meet their expectations. More precisely, disappointed expectations of previous reviewers are adopted by current reviewers and forwarded to future reviewers, leading to lasting negative ratings and an overall suppressed long-term average rating. Thus, a lasting negative effect based on initially stimulated, overshot expectations (of early adopters) results. This lasting effect is also supported by works in the related research field of electronic word-of-mouth (eWOM). Here, Hornik et al. (2015) identified two negativity biases in subsequent eWOM (i.e., eWOM which refers to previous eWOM) that are based on Dynamic Social Impact Theory (Latané and Bourgeois, 2001) and rumor diffusion (DiFonzo and Bordia, 2007). They found that negative as well as positive eWOM leads to more negative than positive eWOM. Further, the negative eWOM is disseminated for a longer period of time than the positive eWOM. The theoretical explanations for these effects are "malicious joy", which leads to more negative subsequent eWOM based on negative eWOM and "jealousy", which leads to more negative subsequent eWOM based on positive eWOM. Both negativity biases are in line with our results. In contrast, items without strong initial rating dynamics avoid such heavy disconfirmation and thus also elude negativity biases and unrealistic lasting expectations which results in a higher long-term average rating for these items. In total, this finding shows that initial rating dynamics do not vanish after the short term but influence the long-term average rating and assessment of items.

These two key findings also have major implications for practice. To begin with, the revealed U-shaped relationship between sequential order and online consumer ratings indicates strong rating dynamics which vary with respect to the order position. Thus, consumers and businesses which strive towards capturing accurate and representative consumer feedback need to take the order position into account when considering a rating, as it is impacted by two disconfirmations. Similarly, review platforms that are willing to provide helpful information to their users should take these dynamics into account (e.g., by not penalizing items which are hit by a strong negative first disconfirmation before the positive second disconfirmation kicks in). Moreover, as a strongly decreasing initial rating dynamic has a negative impact on the long-term average rating, businesses should not encourage early adopters to leave very high initial ratings for an item. Similarly, early adopters themselves should aim to submit an objective and factual rating to not provoke false expectations and disconfirmation by other consumers. Indeed, our second finding shows that this tends to lead to a lower average rating in the long term, which may impede potential consumers from choosing this item any further. Thus, consumers should keep this impact in mind and not focus too strongly on initial one-sided (subjective) ratings. Further, this finding also shows that the long-term average rating – although many consumers pay high attention to it – is not always an appropriate criterion for assessing items, as it

may be negatively impacted by a strong initial rating dynamic. Therefore, platforms could try to delay the publication of initial ratings of early adopters and instead initially support reviews from more neutral consumers. This way, very strong initial in- or decreasing rating dynamics which lead to a negatively biased long-term average rating can be avoided. Another option to reduce this effect would be to make visitors of the platform attentive of strong initial rating dynamics and their impact. Thereby, improved decision support could be established.

6 Conclusion, Limitations and Future Work

Online consumer ratings comprise essential purchase information that influence other potential consumers and provide direct feedback for businesses. These ratings capture consumer assessments at different points in time and in (sequential) order. While extant literature has recognized the importance of sequential and temporal dynamics of consumer ratings, it focuses on the short term and lacks analyses of long-term dynamics. However, this is particularly important, as many items receive ratings over the long term, but existing findings cannot explain these long-term dynamics. In this paper, we examine long-term sequential and temporal dynamics and in particular the impact of initial rating dynamics on the long-term average rating of items. To do so, we apply regression models to an extensive long-term review dataset, spanning 13 years with more than 1.9 million restaurant reviews and a median of more than 350 ratings per item. Our analysis yields two novel key findings for long-term dynamics in online consumer ratings: First, we find a novel long-term sequentially increasing rating trend which leads to a U-shaped relationship between ratings and their order, indicating a second disconfirmation. Second, we reveal that initial rating dynamics have significant impact on long-term average ratings, indicating in particular a lasting effect based on initially stimulated expectations. Thus, "initially controversial" items achieve a lower long-term average rating. Both findings extend the existing body of knowledge and thus allows a deeper understanding of online consumer ratings. Furthermore, they have crucial implications for consumers, businesses and review platforms. First, to capture accurate consumer feedback, it is necessary to take the order position of ratings into account, as the rating is significantly impacted by two disconfirmations. Second, consumers, businesses and platforms should be aware that the long-term average rating is negatively impacted by strong initial rating dynamics. To avoid this effect, early adopters should aim to submit an objective and factual rating while platforms need to provide additional information to support decisions. Moreover, this finding indicates that the long-term average rating is not always an appropriate criterion for assessing items.

Nevertheless, the work at hand has some limitations which could be a starting point for further research. As we used a review dataset of the restaurant domain, further analyses on other datasets of different domains should be conducted to assess the generality of our findings. Moreover, we only focused on the rating part of reviews. Therefore, it would be interesting to analyze whether similar sequential and temporal dynamics also exist in the fine-grained sentiments contained in the textual part of consumer reviews. Despite these limitations and directions for future research, we hope that our work will open doors for further discussions in this exciting area.

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle". In B. N. Petrov and F. Csaki (eds.) *2nd International Symposium on Information Theory*, pp. 267–281. Budapest, Hungary: Akadémiai Kiadó.

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

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

Bjering, E., L. J. Havro and O. Moen (2015). "An Empirical Investigation of Self-Selection Bias and Factors Influencing Review Helpfulness" *International Journal of Business and Management* 10 (7), 16–30.

Burnham, K. P. and D. R. Anderson (2004). "Multimodel Inference: Understanding AIC and BIC in Model Selection" *Sociological Methods & Research* 33 (2), 261–304.

Chevalier, J. A. and D. Mayzlin (2006). "The Effect of Word of Mouth on Sales: Online Book Reviews" *Journal of Marketing Research* 43 (3), 345–354.

Dean, A., D. Voss and D. Draguljić (2017). "Polynomial Regression". In *Design and Analysis of Experiments*, pp. 249–284. Cham: Springer.

DiFonzo, N. and P. Bordia (2007). *Rumor psychology: Social and organizational approaches*. Washington: American Psychological Association.

Godes, D. and J. C. Silva (2012). "Sequential and Temporal Dynamics of Online Opinion" *Marketing Science* 31 (3), 448–473.

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

Ha, S. H., S. Y. Bae and L. K. Son (2015). "Impact of online consumer reviews on product sales: Quantitative analysis of the source effect" *Applied Mathematics & Information Science* 9 (2L), 373–387.

Helversen, B. von, K. Abramczuk, W. Kopeć and R. Nielek (2018). "Influence of consumer reviews on online purchasing decisions in older and younger adults" *Decision Support Systems* 113, 1–10.

Ho, Y.-C., J. Wu and Y. Tan (2017). "Disconfirmation Effect on Online Rating Behavior: A Structural Model" *Information Systems Research* 28 (3), 626–642.

Hornik, J., R. S. Satchi, L. Cesareo and A. Pastore (2015). "Information dissemination via electronic word-of-mouth: Good news travels fast, bad news travels faster!" *Computers in Human Behavior* 45, 273–280.

Hu, N., P. A. Pavlou and J. Zhang (2017). "On Self-Selection Biases in Online Product Reviews" *MIS Quarterly* 41 (2), 449–471.

Kuha, J. (2004). "AIC and BIC: Comparisons of Assumptions and Performance" *Sociological Methods & Research* 33 (2), 188–229.

Latané, B. and M. J. Bourgeois (2001). "Dynamic Social Impact and the Consolidation, Clustering, Correlation, and Continuing Diversity of Culture". In M. A. Hogg and R. S. Tindale (eds.) *Blackwell Handbook of Social Psychology: Group Processes*, pp. 235–258. Oxford, UK: Blackwell Publishers Ltd.

Lederrey, G. and R. West (2018). "When sheep shop: measuring herding effects in product ratings with natural experiments". In: *Proceedings of the 2018 World Wide Web Conference*. Ed. by P.-A. Champin, F. Gandon, M. Lalmas, P. G. Ipeirotis. New York, NY, USA: ACM, pp. 793–802.

Levy, Y. and T. J. Ellis (2006). "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research" *Informing Science: The International Journal of an Emerging Transdiscipline* 9, 181–212.

Li, H., Z. Zhang, F. Meng and Z. Zhang (2019). ""When you write review" matters: The interactive effect of prior online reviews and review temporal distance on consumers' restaurant evaluation" *International Journal of Contemporary Hospitality Management* 31 (3), 1273–1291.

Li, X. and L. M. Hitt (2008). "Self-Selection and Information Role of Online Product Reviews" *Information Systems Research* 19 (4), 456–474.

McCullagh, P. (1980). "Regression Models for Ordinal Data" *Journal of the Royal Statistical Society: Series B (Methodological)* 42 (2), 109–127.

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

Muggeo, V. M. R. (2003). "Estimating regression models with unknown break-points" *Statistics in medicine* 22 (19), 3055–3071.

Muggeo, V. M. R. (2008). "Segmented: An R package to fit regression models with broken-line relationships" *R news* 8 (1), 20–25.

Murphy, R. (2020). *Local Consumer Review Survey 2020*. URL: https://www.brightlocal.com/research/local-consumer-review-survey/# (visited on 03/27/2022).

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

Oliver, R. L. (2014). *Satisfaction: A Behavioral Perspective on the Consumer.* 2nd edition. New York: Routledge.

Park, S., W. Shin and J. Xie (2021). "The Fateful First Consumer Review" *Marketing Science* 40 (3), 481–507.

Raftery, A. E. (1995). "Bayesian Model Selection in Social Research" *Sociological Methodology* 25, 111–163.

Rawlings, J. O., S. G. Pantula and D. A. Dickey (1998). *Applied Regression Analysis: A Research tool.* 2nd edition. New York: Springer New York.

Ricci, F., L. Rokach and B. Shapira (2015). "Recommender Systems: Introduction and Challenges". In F. Ricci, L. Rokach and B. Shapira (eds.) *Recommender Systems Handbook*, pp. 1–34. Boston, MA: Springer.

Schwarz, G. (1978). "Estimating the Dimension of a Model" *The Annals of Statistics* 6 (2), 461–464.

Simonsohn, U. (2018). "Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships With Quadratic Regressions" *Advances in Methods and Practices in Psychological Science* 1, 538–555.

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

Wherry, R. J. (1931). "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation" *The Annals of Mathematical Statistics* 2 (4), 440–457.

Wu, F. and B. A. Huberman (2008). "How Public Opinion Forms". In C. Papadimitriou and S. Zhang (eds.) *Internet and Network Economics*. *WINE 2008. Lecture Notes in Computer Science*, pp. 334–341: Springer, Berlin, Heidelberg.

Wu, F. and B. A. Huberman (2010). "Opinion formation under costly expression" *ACM Transactions on Intelligent Systems and Technology* 1 (1), 1–13.

Yin, P. and X. Fan (2001). "Estimating R2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods" *The Journal of Experimental Education* 69 (2), 203–224.

2.2 Paper: The Way to the Stars: Explaining Star Ratings in Online Consumer Reviews

Current Status	Citation
This paper is **under review in revision** for publication in the journal *Decision Support Systems*.	Binder, M., B. Heinrich, M. Hopf and M. Szubartowicz (2022b). "The Way to the Stars: Explaining Star Ratings in Online Consumer Reviews" Working Paper, University of Regensburg.

The Way to the Stars: Explaining Star Ratings in Online Consumer Reviews

Markus Binder, University of Regensburg, Regensburg, Germany, markus1.binder@ur.de
Bernd Heinrich, University of Regensburg, Regensburg, Germany, bernd.heinrich@ur.de
Marcus Hopf, University of Regensburg, Regensburg, Germany, marcus.hopf@ur.de
Michael Szubartowicz, University of Regensburg, Regensburg, Germany, michael.szubartowicz@ur.de

Abstract: Online consumer reviews are important performance indicators for businesses since they constitute essential sources of information for consumers. To gain detailed insights from these reviews, researchers have already used features (such as the feature *food quality* of a restaurant being part of the general feature perspective *item aspects*) derived from review texts to explain associated star ratings. However, existing literature analyzes only certain feature perspectives, enabling just a partial view. Therefore, we leverage four different feature perspectives expressed in consumer reviews (each comprising easy-to-interpret features) in an explanatory model to study whether star ratings can be explained by these feature perspectives. The evaluation on three large real-world datasets shows that the proposed feature perspectives explain star ratings considerably well (Nagelkerke pseudo R-squared of 65-70%) with substantial contributions of each feature perspective. In particular, the perspective *user characteristics* – rarely discussed in related literature – yields the second highest contribution, while *item aspects* contribute the most. Besides valuable implications for research, our work indeed allows well-founded actions for consumers, web portals and businesses.

1 Introduction

With the growing number of people seeking and purchasing goods online [1], the volume and variety of online consumer reviews on web portals such as Amazon or TripAdvisor are vastly increasing [2–4]. Thereby, online consumer reviews constitute a vital object of study of electronic word-of-mouth (EWOM), which is a major and highly attractive research topic in the field of information systems [5]. Further, it is widely recognized that comprehensible and trustworthy product reviews are a major purchase influence factor [6–8]. Thus, online consumer reviews regarding items (e.g., laptops or

restaurant visits) are important instruments for users of web portals to overcome information asymmetries about these items [9]. In addition, online consumer reviews (as part of EWOM) are important performance indicators for businesses and web portal providers [5]. For instance, careful improvements of products as well as the creation of ideas for new products based on users' preferences and valenced review statements are possible [10,11]. This is, such reviews comprise rich information [12–14] and typically consist of a star rating (e.g., 1 to 5 stars) representing the overall user assessment and a textual part. Besides the frequently analyzed EWOM-dimensions valence and volume, recent research started to investigate the semantic and lexical content of EWOM [5]. Here, a major goal of research is to derive insights from star ratings and the semantic and lexical content of review texts [15,16] in order to understand why users rated an item the way they did. Therefore, it is important to leverage features (e.g., the feature *food quality* of a restaurant) derived from review texts to explain associated star ratings. More precisely, we aim to analyze relationships between several features of different feature perspectives (e.g., the feature perspective *item aspects* including the feature *food quality*) expressed in reviews as independent variables and star ratings as dependent variable, which is enabled by an *explanatory* model [17]. Thereby, it is vital to utilize independent variables representing *easy-to-interpret* features (e.g., features that can be traced back to its semantically related feature terms in the review texts) in such an explanatory model which enables to derive both comprehensible and well-founded insights.

The relevance of such explanatory analyses has been acknowledged by recent works (e.g., [12,18–20]) proposing explanatory models for star ratings based on selected single features. For instance, some works focus on features towards particular *item aspects* in review texts (e.g., food quality of a restaurant) [12,21], while other works aim at specific *user context* features (e.g., dining companions) [22]. Here, existing approaches consider selected features of at most two different feature perspectives in their analyses, enabling only a partial view. For instance, the feature perspective *user characteristics* encompasses personal factors such as user personality or social identity. While it has recently been utilized in the research on personality-based recommender systems (e.g., [18]), it is rarely analyzed in the context of explaining star ratings. In addition, existing

works do not investigate the contribution of each individual feature perspective to the explanatory power of their proposed models. This would give important insights, such as that *user characteristics* – rarely discussed in related literature – constitute the feature perspective with the second highest contribution to the explanation of star ratings, which calls for researchers to incorporate this feature perspective in their analysis of online consumer reviews.

In this paper we are the first (A) to integrate the features of more than two feature perspectives into a unified model for explaining star ratings and (B) to analyze the relative importance of each feature perspective for the explanatory power of this model. Therefore, this work could serve as a first step for enabling a thorough understanding of star ratings (cf. discussion in Section 2.1). With this in mind, we focus on the following research questions:

RQ1: *To what extent can features of different feature perspectives explain star ratings in online consumer reviews?*

RQ2: *How much does each individual feature perspective contribute to the explanatory power of the unified model?*

To address these questions, we derive the four object and person-centered feature perspectives *item characteristics, item aspects, user characteristics* and *user contexts* from the popular *Multiple Pathway Anchoring and Adjustment* (MPAA) model and unify these feature perspectives into one single model used to explaining star ratings. To extract easy-to-interpret features from a very large number of review texts and thus operationalize the feature perspectives, we apply the state-of-the-art deep learning language model BERT [19]. Given this set of extracted features of different perspectives as independent variables, we evaluate their explanatory power by using the generalized ordered probit regression model (GOPM, cf. [20]). First, we find that the proposed feature perspectives explain star ratings considerably well, which is indicated by Nagelkerke pseudo R-squared values of 65% up to 70%. In comparison, similar works reach Nagelkerke pseudo R-squared values of up to 44% or analyze only 1 star and 5 star ratings. Second, calculating partial R-squared values shows substantial contributions of each feature perspective to the explanatory power of the unified model. In particular, *user characteristics* – rarely discussed in

related literature – constitute the feature perspective with the second highest contribution to the explanatory power, while *item aspects* – capturing the user's experience of an item – contribute the most. Additionally, the feature perspective *item characteristics* is able to explain ratings better than *user contexts* for search goods such as laptops while the opposite holds in the restaurant domain.

Our work has several implications on research and practice. First, from a scientific point of view, we enhance the existing body of knowledge in the field of EWOM [5] by our analysis, which poses a first step towards a comprehensive theoretical foundation for the explanation of star ratings in online consumer reviews based on easy-to-interpret features and feature perspectives. Further, our analysis regarding the contributions of each feature perspective to the explanatory power can encourage researchers in the field of EWOM to utilize in particular the feature perspective *user characteristics* in their analyses, which is widely ignored by existing works. Second, the analysis of different feature and feature perspectives allows cross-domain insights (e.g., agreeable users give more positive ratings) and domain-specific insights (e.g., brand loyalty influences laptop ratings). Third, the proposed explanatory model enables a detailed analysis of, for instance, reviews regarding different star rating levels. That is, features important for explaining fine-grained differentiations between individual rating levels (e.g., 4 star and 5 star ratings) can be analyzed in an overall explanatory model. Fourth, from a practical viewpoint, the explanations derived by our model can support web portals and businesses to automatically identify important features that highly influence user assessments (i.e., features with high regression coefficients). By analyzing these important features in detail, existing products can be carefully improved or even new ideas for products can be created. Fifth, these important features support web portals in summarizing user experiences, designing structured multi-criteria rating systems [21] or indicating why users rated an item the way they did. In this way, web portals can ensure that these highly relevant features are easily accessible to consumers when forming attitudes towards items in their purchase decisions (e.g., by providing a structured summary of user experiences for each highly relevant feature).

The remainder of the paper is structured as follows. In the next section, we introduce the theoretical foundations for analyzing consumer reviews, discuss the related work and present the

research gap. In the third section, we derive four feature perspectives from the literature for explaining star ratings and formulate two detailed research questions. Thereafter, in Section 4, we evaluate the explanatory power using three large real-world datasets from different domains (i.e., restaurants, movies and laptops) and present the results. In the subsequent sections, we discuss the evaluation results and outline the implications of the results for research and practice. Finally, we conclude with a summary of the main findings, reflect upon limitations and provide an outlook for future research.

2 Background

In this section, we first present the theoretical foundations for our research. Then, we give an overview of existing works which aim at explaining the users' star ratings and establish the research gap.

2.1 Theoretical Foundations

Many works that discuss and analyze online consumer reviews are based on the notion that such reviews constitute a textual and numerical representation of a user's multiple attitudes towards an item (e.g., [22,23]). Focusing on such user attitudes, the popular *Multiple Pathway Anchoring and Adjustment* (*MPAA*) model by Cohen and Reed [24] constitutes a recognized theoretical foundation. More precisely, the MPAA model incorporates prior research on the formation, recruitment and retrieval of attitudes as well as attitude-behavior relationships into an integrative model. In particular, the literature on attitude representation suggests a relationship between formed attitudes and the behavior of users (e.g., assessing features in a review) [25,26]. Consequently, the MPAA model can provide a foundation to investigate the semantic and lexical content of review texts.

In more detail, Cohen and Reed [24] lay out the body of knowledge supporting the existence of multiple attitudes towards the same item (cf., e.g., [27]). For instance, a person might form an attitude towards a sports car based on its (object-centered) features like acceleration and price. Moreover, a different attitude based on personal values might be formed after the person learns about

the social status accompanied by this car. In order to incorporate the coexistence of multiple (possibly opposed) attitudes, the MPAA model proposes the idea of multiple pathways which lead to the formation of such attitudes. These pathways are categorized into *object-centered* (or *outside-in*) and *person-centered* (or *inside-out*) pathways. Object-centered pathways focus on attitudes which are generated through an actual experience with an object as well as through analytical, combinatorial or analogical cognitive processes. Person-centered pathways involve attitude formation by using the personal value system, social identity or frame of reference. Taken together, these pathways lay the foundations for multiple feature perspectives which enable a differentiated view when explaining star ratings.

Object-centered pathways consider item assessments which are based on the user's already existing attitudes towards certain *item characteristics* or on the user's actual experience of an item through its *item aspects* [24]. In contrast, person-centered pathways are provided by *user characteristics*, such as the user's personality or social identity, and specific situational *user contexts* that influence the user's assessment of an item [23]. Therefore, the object-centered feature perspectives *item characteristics* and *item aspects* as well as the person-centered feature perspectives *user characteristics* and *user contexts* are described in the following:

Based on *item characteristics* like the *genre* of a movie or the *cuisine* of a restaurant, the user's preliminary attitudes and preferences can be analyzed. In particular, the preferences of a user can be determined based on the characteristics of the items the user liked or disliked in the past [28]. That is, even for items unfamiliar to the user, it is aimed to infer preferences based on familiar items with similar item characteristics [24]. Based on item characteristics, a user explicitly builds her or his preconception for an item ex ante. For example, in the domain of restaurants, a user can have an already existing positive attitude towards the value *Thai food* of the item characteristic *cuisine* of a restaurant. In that case, the user's star rating for this restaurant thereby may be influenced through her or his positive attitude towards *Thai food*. This suggests that item characteristics are related to the users' star ratings [29].

In contrast to item characteristics, *item aspects* (e.g., *service* at a restaurant) and their sentiments (often called aspect-based sentiments) capture attitudes, which are formed after actually experiencing an item and not beforehand (cf. [24]). In particular, a user can determine her or his sentiments towards an item aspect in a very detailed way, as the actual experience enables the user to substantiate or modify her or his existing attitudes towards an item's aspects. This experience may also lead to the formation of attitudes towards hitherto undiscovered item aspects. For instance, a user might expect a pleasant *service* before going to a restaurant. After visiting this restaurant and being served by an impatient waiter, the user would form a negative sentiment towards the experienced *service*. In consequence, a negative sentiment could have a high impact on the assigned star rating for this restaurant. Since this perspective comprises detailed user assessments, it is frequently used to analyze and explain star ratings (e.g., [4,12,22,30]).

In contrast to the object-centered perspectives above, *user characteristics* outline personal factors such as user personality or social identity. By definition, user personality aims to capture psychological traits, which account for individual differences in behavior and experience. Amongst other models such as the Myers-Briggs Type Indicator [31], the Five-Factor Model [32] is the most dominant and widely applied personality model comprising the five factors *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness* and *emotional stability*. This model aims to enable a comprehensive, but nonetheless detailed conception of the personality of an individual person. Thereby, the Five-Factor model is referred to as the most comprehensive and parsimonious model of personality [33]. The underlying intuition suggests that the facets of the user's personality allow for a more profound understanding of the user actions, reactions and assessments. Here, studies have shown that the Five-Factor model is particularly useful to examine online behavior in the context of EWOM [33,34]. One reason for that is that (Five-Factor) personality traits effect how individuals attain gratification, for which (the creation of) EWOM is a relevant medium [34]. For example, Hu and Pu [35] discovered that users who score high on *agreeableness* would tend to give higher star ratings. In that line, agreeable users might value harmony and fairness and thus be more inclined not to give extremely negative ratings [24]. Moreover, the analysis in [36] shows, inter alia, that reviews

from users scoring high in emotional stability affect similar users while reviews from users with high emotional range do not affect users with similar personality. Analogously to user personality, the relation of a user's social identity to an item can be a significant influence factor for star ratings. According to [24], social identity can be defined by social categories such as demographics, social roles and shared consumption patterns. Here, users with similar demographic background are expected to rate items similarly. For example, user characteristics such as age or gender might influence a user's star rating for a movie. In total, this indicates that user characteristics can be important to explain star ratings of online consumer reviews.

A further person-centered perspective is the *context of a user*, which, in contrast to item aspects, is not directly related to the rated item. Instead, the user contexts refer to the situational circumstances in which a user interacts with the rated item. Contextual features are, for instance, *time*, *location*, *weather and temperature*, *mood*, and *social encounters* [37]. These user contexts already have been discussed as a potential influencing factor for star ratings [4]. For example, Radojevic et al. [29] analyzed that business travelers tend to be more critical in their ratings of items, which might be reasoned with a higher level of stress on business trips. This indicates that the user contexts can influence star ratings.

2.2 Related Work

In this section, we embed our research into the field of EWOM and discuss existing research, which aims at *explaining* star ratings of online consumer reviews. Regarding the framework of existing EWOM literature by [5], our research can be classified as *evaluation of EWOM* focusing on the *investigation of the semantic and lexical content* of online consumer reviews. In contrast to several existing works in this research strand, which investigate the *coherence* between feature assessments and product-level assessments (e.g., [38]), we analyze the *relation* between features of different perspectives and product-level user assessments. Before we discuss the works in our research strand, we outline and delimit related research strands. Existing works, which focus exclusively on a predictive analysis (e.g., recommender systems) such as [18], [39], [40] or [41] and which do not aim

to explain or interpret the star ratings, are out of scope for our research. As outlined by [17], prediction and explanation are two different objectives and thus need to be assessed differently. When predicting the relation between different variables, the underlying (theoretical) construct is not focused on. In that line, variables used in predictive approaches such as latent factors are not necessarily interpretable and are not aimed to explain the underlying construct. Similarly, works in literature exist, which rely on research techniques such as consumer surveys or group interviews. A restriction of these works is often the limited size of data used for evaluation (usually well below 1,000 observations). As a result, a more complex explanatory model cannot be evaluated on such a smaller dataset, since the resulting ratio of observations to variables in the model would be too small to obtain reliable results [42]. Furthermore, the observations and data used in these evaluations is often influenced by the fact that interviewed users answer the survey solely based on their imagination and expectations but not on, for instance, real experiences, as they actually did not buy, consume nor use an item in reality. Because of these important differences, these works are also out of the scope for our research.

In accordance with the guidelines of standard approaches to prepare the related work (e.g., [43]), we searched the databases ACM Digital Library, AIS Library, EBSCO Host, IEEE Xplore and ScienceDirect without posing a temporal restriction using the search term *(explain* OR explan* OR understand*) AND ("star ratings" OR "consumer ratings" OR "user ratings" OR "customer ratings") AND review**. This search led to 305 papers, which were manually screened based on title, abstract and keywords resulting in 14 papers (the vast majority of the 305 papers focused on predictive analyses or analyzed the helpfulness of online consumer reviews for other users). A detailed analysis of these 14 papers led to 10 papers relevant for our research. Additionally, we performed a forward and backward search starting from these 10 relevant papers. After all, 17 papers were identified as relevant for our work at hand and are grouped by their considered feature perspective(s) in Table 1. These works are discussed in the following regarding (A) the considered feature perspectives and (B) the assessment of contributions of the considered feature perspectives (i.e., the relative importance of the feature perspectives) to the explanatory power of the proposed models.

	Ad (A)				**Ad (B)**
	Object-centered Feature Perspectives		*Person-centered Feature Perspectives*		*Assessing the contributions of feature perspectives to the explanatory power of the model*
	Item Characteristics	*Item Aspects*	*User Characteristics*	*User Contexts*	
[12]; [22]; [44]; [20]; [45]; [46]; [47]; [48]	n/a	Selected features such as food and price for restaurants or cleanliness and service	n/a	n/a	n/a
[29]	Selected features such as hotel's star classification and hotel price	n/a	n/a	Only the features trip purpose and date	n/a
[4]; [49]; [50]; [51]; [52]; [53]; [54]	n/a	Selected features such as food, service and price	n/a	Selected features such as trip purpose and travel party	n/a
[55]	n/a	n/a	Selected features such as user personality and metadata	Only the feature trip purpose	n/a

Table 1. Existing Approaches for Explaining the Star Ratings of Online Consumer Reviews

Ad (A): The first set of works contains eight approaches considering only item aspects. Binder et al. [20] aim at a methodological contribution by proposing the GOPM to analyze star ratings and evaluate this model against the common linear regression model. To this end, aspect-based sentiments are only used for demonstration purposes. In their analysis, Jabr et al. [12] focus on 1 star and 5 star ratings aiming to retrieve unambiguous sentiment data, which concentrates their results on explaining the basic rating tendency. Moreover, Chatterjee [22], Chen et al. [44], Guo et al. [47], Linshi [45] and Liu et al. [48] also aim to explain star ratings based on aspect-based sentiments, but do not provide a detailed analysis regarding different steps of the rating scale, which may be interesting in their research. Lastly, Lacic et al. [46] analyze star ratings by determining correlation coefficients between these ratings and individual aspect-based sentiments rather than establishing an explanatory model.

The second set of works comprises only the work of Radojevic et al. [29], which consider single features being part of the perspectives item characteristics and user contexts to explain star ratings. However, these features are extracted only from structured data, excluding the information contained in review texts.

Indeed, there also exist seven approaches that analyze the impact of item aspects combined with user contexts on star ratings. The two consecutive works of [51] and [50] analyze star ratings to

determine how much these ratings vary between reviews for different items and within the same item in their model. Thereby, a different set of coefficients for each item is used, which limits the reliability of the results for items not having a considerably high number of available reviews. Ye et al. [52] aim to explain the sub-ratings for service quality and value for money rather than the overall star rating. The work of [53] provides a detailed explanatory analysis, in particular, of the coefficients in their regression model, but with a special emphasis on the traveling domain. Further, Luo and Tang [49] and Xiang et al. [4] aim to examine the influence of aspects and contexts on the star rating. Another recent work analyzes the impact of the feature perspectives item aspects and user contexts on star ratings in the domain of airline traveling [54]. The authors also utilize user features on a cultural-level, that is, these features are derived solely based on the citizenship of a user. However, the authors state that "people within a same culture can have different types of personality traits, which [...] cannot be measured" by these features and recommend that "future researchers could thereby choose more suitable or alternative measures" for the feature perspective user characteristics. In particular, this means that their considered country-related features are hardly suitable for a review-level analysis of star ratings since all users from the same country have the exact same feature values for this feature perspective. Hence, only suitable features from only two perspectives are considered in this work. In addition, none of these seven works investigates whether their explanatory models can explain different steps of the rating scale (e.g., why users rated an item with 4 or 5 stars).

Finally, there also exists a recent work that analyzes the impact of user characteristics on star ratings [55]. In particular, this work focuses on the impact of the Five-Factor user personality traits on star ratings. Additionally, they analyze one feature ('travel type') as user context in the hotel domain (i.e., business trips vs. leisure trips).

Ad (B): Since the first set of works focuses solely on one feature perspective, an analysis and comparative assessment regarding the contributions of different feature perspectives to the explanatory power of the models is not possible. The sets of works considering two feature perspectives also lack such an analysis, which would give important insights, even though only two

feature perspectives are considered. While the work of [54] investigates interdependent moderator effects between the considered feature perspectives, the contributions of the two feature perspectives to the explanatory power of the model are not assessed in this work either.

To conclude, there are already interesting works that aim to explain the overall star ratings of online consumer reviews based on different features and feature perspectives. However, these contributions (A) only consider selected suitable features (often only one single feature) of at most two feature perspectives in their analysis. Further, (B) none of the existing works assesses the contribution of each feature perspective in terms of explanatory power which could allow for valuable insights on the relative importance of different feature perspectives (and their features) on user assessments in online reviews. In this paper, we aim at filling the identified research gap by (A) leveraging more than two feature perspectives for explaining star ratings and by (B) analyzing the relative importance of each feature perspective for the explanatory power of the proposed model.

3 Explaining Star Ratings in Online Consumer Reviews

To address the identified research gap, we derive four object and person-centered feature perspectives from the MPAA model and unify these feature perspectives into one single model to explain star ratings in online consumer reviews (cf. Figure 1).

Regarding *object-centered feature perspectives*, Cohen and Reed [24] discuss that a user's attitudes depend on both initial knowledge and actual experience. To be more precise, *item characteristics* like the genre of a movie are usually known before experiencing an item and thus can be used to form an initial attitude by evaluating these (known) item characteristics. This corresponds to the pathway *Analytical Attitude Construction* of the MPAA model. Thereby, star ratings may be influenced in a positive or negative way depending on the (expected) peculiarity or importance of item characteristics. For instance, focusing only on item characteristics, a user might form positive attitudes towards a movie because she or he likes the genre and director of the movie, but might establish also some negative attitudes because she or he has no sympathy for the main actor. In addition, actual exposure to an item may influence its assessment. In this case, an attitude is formed

based on a user's directly captured perception of the item, which can be structured by *item aspects*. While most online consumer reviews are composed after a direct experience, the corresponding pathway *Direct/Imagined Experience with the Object* of the MPAA model also encompasses attitudes formed by a simulated experience. Such experiences, however, can be structured according to item aspects as well. Therefore, the detailed analysis regarding item aspects can give further insights into how the user's overall star rating can be explained.

With regard to *person-centered feature perspectives*, Cohen and Reed [24] argue that a user might generate an attitude by relating and evaluating an item to her or his characteristics such as personal traits or social identity. This corresponds to the pathway *Social Identity-Based Attitude Generation* of the MPAA model. As such, *user characteristics* like agreeableness can be examined to point out similarities and differences between individuals which can be reflected in similar or dissimilar star ratings. Additionally, an important factor of the MPAA model is the (temporal) change of attitudes due to contextual variations. Depending on a certain context, different subsets of personal beliefs and values might be used to form an attitude. For instance, stress situations and time constraints might lead to favoring junk food over healthier options while other contexts might have the reverse effect. In the MPAA model these contextual variations are reflected in the pathway *Value-Driven Attitudes*. As a consequence, user contexts may play an important role for the explanation of star ratings.

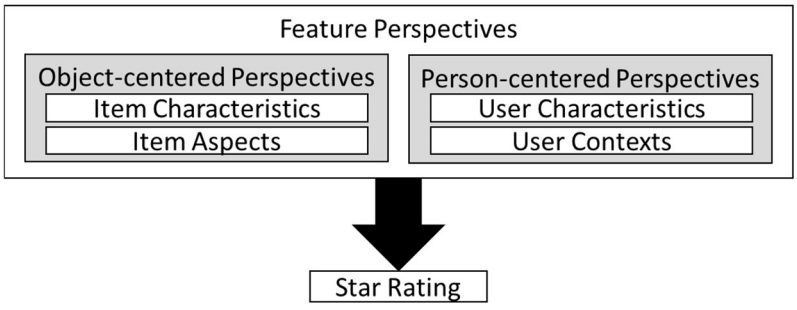

Figure 1. Research Model for Explaining Star Ratings in Online Consumer Reviews

As noted in the previous section, related work focuses on at most two feature perspectives. We argue that a broader view should be established to explain star ratings in online consumer reviews and hence, we pursue RQ1. While we analyze the overall explanatory power in RQ1, it is of

high relevance to analyze how much each individual feature perspective (and their features) contributes to this overall explanatory power. This enables to investigate if one individual feature perspective surpasses all others regarding its contribution to the overall explanatory power or whether the combination of the object-centered and person-centered feature perspectives is of additional value. This would substantiate the theoretical grounding given by the MPAA model and therefore, RQ2 is proposed.

To answer these two research questions, we deliberately outline quantitative analyses instead of following a common hypothesis-driven framework. As we aim to evaluate the quantitative extent of the explanatory power from different angles, a solely hypothesis-driven framework would limit the scope of our analyses. We argue that using the GOPM in combination with the Nagelkerke pseudo R-squared (cf. Section 4.2 below) allows for deeper and more differentiated insights of the results of our analyses. Moreover, it has been recognized that focusing on significance results can be misleading when analyzing very large datasets (such as the review datasets analyzed in our evaluation) [56]. As the significance of an effect does not provide any information about the magnitude of the effect, it is even argued that the "notion of statistical significance is not that relevant to big data" [57]. Thus, we focus on quantitative analyses.

4 Analysis and Results

We start this section by introducing the selected datasets and describing their preparation for our evaluation. Thereafter, we outline the methodology for our explanatory analysis. We end the section by presenting the results for RQ1 and RQ2.

4.1 Datasets and Data Preparation

For our analysis, we used three large real-world review datasets from the commonly utilized review domains of restaurants, movies and laptops. Reviews of these three domains are also used for analyses in related research fields such as sentiment analysis or design of EWOM systems [15,58] and allow for a broad view of different types of (reviews for) products and services. For instance,

restaurants and movies constitute experience goods, which are goods that have to be mainly experienced by the user to properly assess their quality. In contrast, laptops constitute search goods, which are goods whose quality can be assessed to a greater extent without personal experience [59]. By considering these three multi-faceted datasets for the evaluation, it is possible to derive cross-domain as well as domain-specific insights (e.g., features being important in only one domain). While their properties are typical for online consumer reviews, the three datasets exhibit a higher diversity representing three varying market fields of e-commerce.

In more detail: The restaurant dataset consists of 2.4m reviews for restaurants, bars and cafés in New York City from an established web portal for reviews regarding local businesses. The movie dataset consists of 1.2m reviews for movies and other video content (e.g., documentaries, recorded concerts, etc.), while the laptop dataset consists of 270k reviews for laptops, notebooks and tablet computers, both originating from the Amazon review dataset provided by [60]. Thereby, each review of the above datasets consists of a textual consumer review with an associated star rating on a five-tier scale from 1 star to 5 stars. In order to avoid biases due to specific time frames, the datasets contain reviews from large time periods of ten years or more (time span restaurants: 2008-2017; time span movies: 2000-2018; time span laptops: 2002-2018). Moreover, the datasets cover a broad range of items (e.g., from bistros to gourmet restaurants as well as from economical to high-end laptops). Further, each dataset exhibits the widely recognized "J-shaped" rating distribution (e.g., [61]). To avoid biases due to the skewed rating distribution, we used stratified samples of the datasets with equal rating distributions, similar to [62] and [23]. Thereby, each sample is large enough to analyze

Dataset	**Restaurants**	**Movies**	**Laptops**
# of reviews	2,396,650	1,167,071	271,883
Rating distribution, i.e., (relative) # of ratings per level of star rating	1 star: 9% (~207k) 2 stars: 9% (~214k) 3 stars: 16% (~389k) 4 stars: 34% (~807k) 5 stars: 33% (~779k)	1 star: 8% (~98k) 2 stars: 6% (~65k) 3 stars: 10% (~117k) 4 stars: 19% (~227k) 5 stars: 57% (~660k)	1 star: 17% (~45k) 2 stars: 7% (~19k) 3 stars: 8% (~23k) 4 stars: 18% (~50k) 5 stars: 49% (~134k)
# of reviews in the sample with equal rating distribution	500,000 [100,000 per rating level]	250,000 [50,000 per rating level]	75,000 [15,000 per rating level]
# of users in the sample # of items in the sample	233,854 10,480	208,787 13,677	69,091 3,441

Table 2. Description of the Datasets

various independent variables of different feature perspectives in an explanatory model. That is, the number of events per independent variable (EPV) is higher than 1,000 in our analyses as the smallest sample has 75,000 reviews. This is considered as clearly sufficient in literature (e.g., [63]). An overview of the basic specifications regarding the datasets is given in Table 2.

To evaluate our research questions, we operationalized both the person-centered and object-centered feature perspectives (cf. Section 2) as outlined in the following. The features for the perspective *item characteristics* are directly given in each dataset as structured data. In contrast, the features of the feature perspectives *item aspects*, *user characteristics* and *user contexts* are contained in the unstructured review texts. Here, features can be extracted from textual data by using either unsupervised or supervised methods. Unsupervised extraction methods result in abstract representations that are – a priori – independent of any predefined feature or feature perspective. For example, the unsupervised extraction method topic modeling yields topics comprising a specific set of cooccurring words from review texts. Thereby, existing literature (e.g., [64]) argues that it is very challenging to interpret such abstract representations, as it remains unclear what they really mean, and that these abstract representations do not align with predefined features in general. In addition, Vallurupalli et al. [65] argue that the interpretations of topics strongly depend on the human individuals interpreting the topics. Further, they state that the findings obtained from topic modeling is also highly dependent on the used datasets, even if they are of the same domain. Thus, such findings can hardly be generalized to other datasets or domains, as strongly different topics could be identified. In total, these abstract representations require additional work to derive comprehensible insights. Therefore, we decided to choose a supervised feature extraction method based on BERT [19], which is a state-of-the-art deep learning language model. Here, we first selected and analyzed features and the corresponding feature terms in the review texts for each of the three feature perspectives in line with existing works (cf. Table 3). In particular, this initial analysis showed that each such feature can be traced back to semantically related feature terms, which entails a direct interpretation. By individually training the supervised language model BERT on annotated data (i.e., feature terms) for each feature, the extraction of these features is enabled for a large number of

reviews in the considered datasets (cf. Table 2). Doing this, for instance, the term "*bartender*" can be extracted as semantically related feature term for the item aspect *service* in the sentence "*The bartender was charming*". Here, the language model BERT is further able to identify semantic and lexical representations of natural language [66] by considering whole sentences for feature extraction, which enables a semantically sensitive feature extraction. For instance, the word "*bartender*" would be extracted for the item aspect *service* in the sentence "*The bartender was charming*", but not in the sentence "*The mojito was listed as bartender's choice"* due to different semantical meanings of the term *"bartender*". To further improve the quality of the feature extraction (cf. Section 5), we used the post-trained BERT models for each specific domain of our considered datasets [67], since the post-trained BERT models have a stronger alignment to the domain-specific use of language. Summing up, by the use of this supervised deep learning language model, we are able to extract and utilize features, which are easy-to-interpret (due to its semantically related feature terms in the review texts).

An overview of the considered features for each feature perspective and each dataset is given in Table 3. We selected five *item characteristics* for each dataset with the lowest pairwise correlations and a sufficient number of occurrences (i.e., assigned to more than 10% of items), such as movie genre or cuisine, which is in line with approaches such as [68]. Furthermore, for each dataset we extracted six item aspects, five user characteristics and five user contexts (cf. Table 3). To capture both the users' situational circumstances and actual experiences of items expressed in the review texts, we extracted the sentiments towards features of the perspectives *user contexts* and *item aspects* [50]. More precisely, we extracted *item aspect*-based sentiments and *user context*-based sentiments from the review texts by firstly conducting aspect term and context term extraction and subsequently conducting the task of term-based sentiment classification. For instance, in the exemplary sentence "*The waiter was very friendly.*" first the aspect term "*waiter*" was extracted by BERT and assigned to the item aspect *service*. Subsequently, BERT assigned a positive sentiment towards that aspect term based on the term "*very friendly*". Here, all extracted terms which could not

be assigned to a specific item aspect or user context were subsumed under the features *miscellaneous item aspects* or *miscellaneous user contexts*.

Dataset	**Restaurants**	**Movies**	**Laptops**
Considered item characteristics (independent variables $x_{IC1}, \dots, x_{IC5}$; short x_{IC})	5 characteristics (in line with [73]): cuisine, happy hour specials, noise level, private parking lot, vegetarian food	5 characteristics (in line with [74]): director, price level, genre, languages, cast	5 characteristics (in line with [74]): brand, graphic card, hard drive, processor, memory
Considered item aspects (independent variables $x_{AS1}, \dots, x_{AS}$; short x_{AS})	(Sentiments towards) 6 aspects (in line with [20]): ambience, food quality, food quantity, price, service, miscellaneous	(Sentiments towards) 6 aspects (in line with [75]): acting, story, cinematography, price, music, miscellaneous	(Sentiments towards) 6 aspects (in line with [76]): battery, performance, price, screen/design, support, miscellaneous
Considered user characteristics (independent variables $x_{UC1}, \dots, x_{UC5}$; short x_{UC})	5 characteristics of the Five-Factor Model (in line with [77]): extraversion, emotional stability, agreeableness, conscientiousness, openness to experience		
Considered user contexts (independent variables $x_{UCxt1}, \dots, x_{UCxt5}$; short x_{UCxt})	5 context variables (in line with [37]): location, time, social, weather, miscellaneous	5 context variables (in line with [78]): purchase type, intended use, social, time, miscellaneous	5 context variables (in line with [78]): intended use, operating system, software, connectivity, miscellaneous
Multicollinearity between the independent variables	Average VIF: 1.28 Maximum VIF: 2.59	Average VIF: 1.38 Maximum VIF: 2.78	Average VIF: 1.23 Maximum VIF: 1.76

Table 3. Features of the Datasets after Data Preparation

Before evaluating our model in detail, we analyzed the quality of the preliminary analysis, which means, the aspect term extraction conducted by the deep learning language model. Thereby, F1 scores of 0.78, 0.75 and 0.76 for restaurants, movies and laptops were achieved. Based on this extraction, the aspect term-based sentiment classification yielded F1 scores of 0.84, 0.86 and 0.80, respectively. All F1 scores are comparable to the state-of-the-art [67]. In particular, to extract *user characteristics*, we also trained an individual BERT model for each of the Five-Factor personality traits using a common essay dataset containing personality annotations [69], while structured data regarding the users' social identity (e.g., with respect to demographics) was not available in the datasets. The average accuracy of the resulting BERT personality models was 58%, which coincides with the state-of-the-art validity for Five-Factor personality detection from text on the standard essay benchmark dataset [70]. Moreover, test-retest correlations on the consumer review datasets for

successive 6-month intervals were 0.73 on average. Thus, the reliability of the applied BERT personality models is in line with Five-Factor personality detection based on questionnaires (with test-retest correlations typically ranging from 0.65 to 0.85) and similar to existing approaches extracting Five-Factor personality traits from social media texts [71].

Finally, to verify the stability of our explanatory model, multicollinearity between the independent variables was measured by the variance inflation factor (VIF). The maximum VIFs ranged from 1.76 to 2.78 and the average VIFs ranged from 1.23 to 1.38, whereby VIF values less than ten are uncritical regarding model stability [72].

4.2 Methodology

As introduced above, our explanatory model comprises the feature perspectives item characteristics (IC), item aspects (IA), user characteristics (UC) and user contexts (UCxt). To explain star ratings and evaluate the explanatory power, we use the GOPM and the Nagelkerke pseudo R-squared [79] both as proposed by [20]. The methodological reasons for this choice are outlined in the following. In general, the GOPM is based on the classical ordered probit model [80]. According to the classical ordered probit model, underlying linear preferences $R_*^i \in \mathbb{R}$ are modelled using the independent variables x_{IC}, x_{IA}, x_{UC} and x_{UCxt} representing the four feature perspectives (cf. Table 3). To ensure reliable results (indicated by a high EPV value, cf. Section 4.1), the same set of coefficients for all reviews is used. This leads to a preference model given by

$$R_*^i = \beta_{IC} x_{IC}^i + \beta_{IA} x_{IA}^i + \beta_{UC} x_{UC}^i + \beta_{UCxt} x_{UCxt}^i + \epsilon, \qquad (1)$$

where $\beta_{IC} (= \beta_{IC1}, \beta_{IC2}, \dots, \beta_{ICn})$, β_{IA} , β_{UC} and β_{UCxt} denote the parameters with respect to the independent variables $x_{IC}^i \big(= x_{IC1}^i, x_{IC2}^i, \dots, x_{ICn}^i\big)$, x_{IA}^i, x_{UC}^i and x_{UCxt}^i in the i-th review, and $\epsilon \sim N(0,1)$ denotes the random error term. Then, a discrete random variable $R^i \in \{1, \dots, 5\}$ and thresholds $\theta_1, .., \theta_4$ are used to estimate the actual star rating $r^i \in \{1, .., 5\}$ in the review i, which means, $R^i = 1$ for $R_*^i \leq \theta_1$, $R^i = 2$ for $\theta_1 < R_*^i \leq \theta_2$, … , $R^i = 5$ for $R_*^i > \theta_4$. That is, the parameters $\beta_{IC}, \beta_{IA}, \beta_{UC}$ and β_{UCxt} as well as the thresholds $\theta_1, \dots, \theta_4$ are determined according to the

classical ordered probit model.

In addition to this classical ordered probit model, the GOPM methodically uses different coefficients $\beta_{IC}^1, .., \beta_{IC}^4$ instead of a fixed coefficient β_{IC} (analogous for the other perspectives) to account for varying impacts of the independent variables over the rating scale. This means, for each independent variable v as well as for each rating step between 1 and 5, we can determine a particular coefficient. More precisely, the GOPM for the evaluation is given by

$$R^i \leq j \quad \text{if} \quad \beta_{IC}^j x_{IC}^i + \beta_{IA}^j x_{IA}^i + \beta_{UC}^j x_{UC}^i + \beta_{UCxt}^j x_{UCxt}^i + \epsilon \leq \theta_j \quad \text{for} \quad j = 1,2,3,4. \tag{2}$$

By assigning preference intervals of different sizes to the star ratings, the GOPM can reflect uneven distances within the rating scale, for instance, in contrast to a common linear regression model. As analyzed by [20], in the restaurant domain a rating level of 4 is far closer to a rating level of 5 with respect to the underlying preference than to a rating level of 3. Further, the GOPM accounts for varying impacts over the rating scale by allowing varying coefficients $\beta_{IC}^1, \beta_{IC}^2, \beta_{IC}^3$ and β_{IC}^4. For instance, an unfriendly waiter in a restaurant (i.e., a negative sentiment towards the aspect *service*) may easily drive a user to assign the lowest star rating, while a pleasant service alone will in general not be sufficient to assign the highest star rating.

To assess the explanatory power of the GOPM for star ratings, we use the Nagelkerke pseudo R-squared. This measure compares the likelihood of the GOPM to a null-model [81], which does not take the independent variables from the four feature perspectives into account. That is, the null-model does not distinguish between different reviews, but still determines the thresholds $\theta_1, .., \theta_4$ according to the rating distribution. In detail, the used comparison of likelihoods is equal to the common R-squared measure in case of a linear regression. However, to account for the transformation on the discrete rating scale (cf. Equation (2)), additionally a rescaling to the range [0,1] is used (as denominator in Equation (3)). Overall, and according to [20], the Nagelkerke pseudo R-squared is given by

$$\mathcal{R}^2_{Nagelkerke} = \frac{1 - \left[\frac{L_{Null-Model}}{L_{GOPM}}\right]^{2/M}}{1 - L_{Null-Mode}{}^{2/M}}, \tag{3}$$

where L_{GOPM} and $L_{Null-Model}$ denote the value of the likelihood function at the maximum likelihood estimate of the GOPM and the null-model, respectively. Further, M denotes the number of observations in the model. Thereby, the range $[0,1]$ of the Nagelkerke pseudo R-squared measure is in accordance with the common R-squared measure for linear regression models. We denote this overall explanatory model, which comprises the GOPM and the feature perspectives item characteristics, item aspects, user characteristics and user contexts, as unified model in the following.

4.3 Results

In the following, we present the results regarding the research questions RQ1, RQ2 based on the three real-world datasets of restaurant reviews, movie reviews and laptop reviews.

Ad RQ1: Overall, our analysis for explaining star ratings yields a Nagelkerke pseudo R-squared value (cf. Equation 3) of 69.8% on the restaurant dataset, 64.9% on the movie dataset and 65.0% on the laptop dataset. For a more detailed analysis, we evaluated how well the unified model explains the star ratings for different steps of the rating scale. To assess the explanatory power for each rating level, we applied Equation 3 separately for each subset of reviews by the assigned star rating. As the results in Table 4 show, star ratings are best explained for reviews with 1 star or 5 star ratings.

Dataset	Rating Levels					Overall
	1 Star Reviews	2 Stars Reviews	3 Stars Reviews	4 Stars Reviews	5 Stars Reviews	
Restaurants (Nagelkerke Pseudo R^2)	81.3 %	62.9 %	51.6 %	64.0 %	80.0 %	69.8%
Movies (Nagelkerke Pseudo R^2)	75.4 %	54.6 %	43.6 %	61.3 %	78.9 %	64.9%
Laptops (Nagelkerke Pseudo R^2)	72.9 %	49.8 %	40.2 %	63.3 %	83.6 %	65.0%

Table 4. Explanatory Power for Different Rating Levels

Further, we also examined the coefficients for the variables of the four feature perspectives (as introduced in Table 3). In detail, we analyzed the coefficients $\beta_v^1, \beta_v^2, \beta_v^3, \beta_v^4$ for each variable v in the model built on each dataset. Here, the coefficients for the variables *weather* in the restaurant domain and *miscellaneous user contexts* in the laptop domain were statistically significant with $p < 10^{-2}$ and all other variables v were statistically significant with $p < 10^{-9}$ (cf. Table 5). Due to

length restrictions, the average coefficients $\overline{\beta_v}$=$(\beta_v^1 + \beta_v^2 + \beta_v^3 + \beta_v^4)/4$ regarding the different rating steps are presented only for selected variables v in Table 5 (different coefficients $\beta_v^1, \beta_v^2, \beta_v^3, \beta_v^4$ always had the same sign). As given in Table 5, for instance, the coefficients for the user characteristic *neuroticism* indicate a negative effect on the star rating of a restaurant, movie or laptop. A positive effect is indicated, for instance, by the coefficient of the user characteristic *agreeableness* across all three domains.

	Independent Variable	Restaurant Coefficient	Movie Coefficient	Laptop Coefficient
Item Aspects	service	0.266***		
	support			0.147***
	price	0.055***	0.036***	0.154***
	food quality	0.583***		
	story		0.626***	
	performance			0.327***
Item Characteristics	vegetarian food	0.018***		
	language		0.065***	
	brand			0.160***
User Characteristics	agreeableness	0.240***	0.459***	0.122***
	neuroticism	-0.234***	-0.232***	-0.344***
	conscientiousness	-0.263***	-0.045***	0.038***
	openness	0.158***	0.055***	-0.022***
	extraversion	0.154***	0.143***	0.021***
User Contexts	location	0.215***		
	purchase type		0.130***	
	intended use			0.073***

$***: \boldsymbol{p < 10^{-9}}$; $**: \boldsymbol{p < 10^{-5}}$; $*: \boldsymbol{p < 10^{-2}}$

Table 5. Selected Coefficients of Easy-to-interpret Features for the Different Domains

Ad RQ2: To analyze how much each feature perspective contributes to the explanatory power, we evaluated partial R-squared values [82]. That is, we determined how much additional explanatory power is gained by adding a single feature perspective (i.e., by comparing to a model consisting of only the other three perspectives). To directly compare the results to the explanatory power (e.g., Nagelkerke R-squared of 69.8% for the restaurant domain), we assessed the contribution of the each feature perspective by scaling the partial Nagelkerke R-squared values to this benchmark (e.g., cf. [83]).

Dataset	Feature Perspective			
	Item Characteristics	Item Aspects	User Characteristics	User Contexts
Restaurants (69.8% in sum)	1.7%	49.0%	9.8%	9.3%
Movies (64.9% in sum)	5.1%	39.7%	16.0%	4.1%
Laptops (65.0% in sum)	8.9%	44.7%	9.0%	2.4%

Table 6. Contribution of Each Individual Feature Perspective to the Explanatory Power

The results of this analysis are presented in Table 6. When considering individual feature perspectives, item aspects contribute the most to the explanatory power across all domains in our evaluation, followed by user characteristics. For restaurant reviews, user contexts contribute more than item characteristics, whereas for laptop reviews the contribution of item characteristics is higher in comparison.

5 Discussion

In this section, we discuss the above presented results for each research question.

Ad RQ1: There are several reasons indicating that the unified model explains the star ratings of online consumer reviews well across various domains with each domain containing different types of products or services:

1) [HIGHER EXPLANATORY POWER IN RELATION TO SIMILAR WORKS] The explanatory power of the unified model is higher compared to other explanatory models. For instance, the authors of [7], which use the same statistical model (i.e., GOPM), achieve a Nagelkerke pseudo R-squared value of 44% in the restaurant domain with their explanatory model based only on item aspects. Within the same domain, the Nagelkerke pseudo R-squared value of the analysis at hand reaches nearly 70%. Further, the authors of [12] analyze the explanatory power based on datasets restricted to 1 star and 5 star ratings, which contain Amazon reviews for different product categories (e.g., grocery and gourmet food). Using the McFadden pseudo R-squared, Jabr et al. [12] achieved a maximum value of 80% with an average of 64%, whereas our evaluation yields the maximum McFadden pseudo R-squared value of 88% with an average of 86% across our three datasets also evaluated only on 1 star

and 5 star ratings. This is in line with our more detailed analysis regarding different rating levels (cf. Table 4), which supports the expectation that star ratings are best explained for 1 star and 5 star ratings as the associated review texts contain words (e.g., sentiments) that clearly indicate an extreme star rating. The reason for the higher explanatory power is that it comprises features of different and additional object- and person-centered feature perspectives of the MPAA model.

2) [EASY-TO-INTERPRET FEATURES EXTRACTED FROM TEXT BY STATE-OF-THE-ART TECHNIQUES] Most of the existing works extract the features from the review texts. The works [12,44,45,47,49,53] use unsupervised topic modeling approaches for generating abstract representations of reviews resulting in aspects that have to be interpreted manually in a time-consuming manner. The works [4,20,22,48,50,51,55] utilize lexicon-based extraction techniques, which achieve much lower validity for the feature extraction compared to the state-of-the-art feature extraction techniques such as BERT ([19]). Different to all of these works, the works [29,46,52,54] do not analyze the vital information contained in the textual parts of online consumer reviews for feature extraction, but use features that have been queried from the user when making a review. Answering such queries is an additional effort for users. Further, the applicability of analyses regarding such features is limited to specific domains (e.g., airline traveling). In contrast, we used easy-to-interpret features extracted from text by a supervised state-of-the-art deep learning model (cf. Section 4.1). These features ensure that the analyzed feature perspectives are of high validity, directly comprehensible and allow a deeper analysis and meaningful explanations of star ratings, even for different domains. In the following, we analyze and discuss coefficients from the GOPM (Table 5) to illustrate both *cross-domain* and *domain-specific* insights based on these easy-to-interpret features.

2.1) [CROSS-DOMAIN INSIGHTS] Cross-domain insights can be derived from all feature perspectives. For instance, the perspective user characteristics, which has the same features across all domains, enables cross-domain analyses. Here, the results of our evaluation on different domains substantially extends existing insights [55]. First, users with high *agreeableness* tend to give higher star ratings, represented by positive coefficients $\overline{\beta_{aggr.}}$ across all three domains. This indicates that agreeable users behave friendly and generously [35]. In contrast, users with high *neuroticism* might be

oversensitive and easily aggravated by items [32], which reasons its negative impact, represented by negative coefficients $\overline{\beta_{neuro.}}$ across all three domains. These observations extend and generalize the findings of [55], which analyzed star ratings for a single domain of experience goods (i.e., hotels). However, the observation of [55] stating that *openness* has a positive effect on star ratings does not hold true in general. Our results show that openness indeed typically has a positive effect on star ratings for experience goods (e.g., restaurants or movies), but has a negative impact for the search good laptops as represented by the coefficients $\overline{\beta_{openn.}}$ in Table 5. As users with high scores on openness like novelty and are enterprising [85], they seek new experiences which can easily be found by testing new foods or movies. Conversely, searching laptops often involves comparing technical details in specifications and data sheets, which is usually not an inspiring experience thus resulting in negative impact. Interestingly, *conscientiousness* has the opposite effect. This might be due to the fact, that conscientious users tend to be well prepared and informed when purchasing an item, which is easier for search goods (e.g., laptops). Finally, *extraversion* consistently has a positive effect, which is plausible, since extraversion also measures a person's tendency to express positive emotions [35].

2.2) [DOMAIN-SPECIFIC INSIGHTS] Domain-specific insights can be derived from all four feature perspectives. For example, we found that the users' current *location* and proximities to restaurants significantly influence their star ratings, which is represented by $\overline{\beta_{location}} = 0.215$ for user contexts in the restaurant domain. Presumably, users might favor conveniently located restaurants to avoid the time and organization effort to travel to and from the restaurant. Moreover, our results show that the item characteristic *brand* considerably influences a user's star ratings in the laptop domain (e.g., $\overline{\beta_{brand}} = 0.160$). This can be reasoned by the relatively high brand loyalty associated to electronic devices like laptops [86]. When shopping for a laptop, users might use brands to infer the performance or quality of a product. In particular, this indicates the high potential of the proposed explanatory analysis enabling differentiated insights for varying types of services and products (e.g., popular products vs. niche products). For instance, we found in a (first) product-differentiated analysis that the item characteristic *brand* is important for laptops of different vendors showing

robust results. In particular, the importance is even (slightly) higher for vendors like *apple*. Further, the item aspects *food quality* ($\overline{\beta_{food\ qual.}} = 0.583$) and *service* ($\overline{\beta_{service}} = 0.266$), which are experienced at a restaurant, are even of higher importance. That is, if a user does not enjoy the food and the service in a restaurant, she or he will typically assign a lower star rating and vice versa. In contrast, the price range of a restaurant is often known or at least anticipated prior to the visit, which may lead to the comparably lower importance of the item aspect *price* ($\overline{\beta_{price}} = 0.055$). Moreover, by using the GOPM (cf. Section 4.2), we are able to inspect four coefficients (cf. Equation (2)) for each feature. For instance, the coefficients regarding the aspect *food quality* in the restaurant domain are given by $\beta^1_{food\ qual.} = 0.373$, $\beta^2_{food\ qual.} = 0.633$, $\beta^3_{food\ qual.} = 0.761$ and $\beta^4_{food\ qual.} = 0.564$. This indicates that the item aspect *food quality* is comparably more important for users to distinguish 3 from 4 stars ratings ($\beta^3_{food\ qual.} = 0.761$) than 1 star from 2 stars ratings ($\beta^1_{food\ qual.} = 0.373$). These findings further emphasize the high sensitivity of the GOPM for star rating explanations.

3) [PARTIAL VIEW PROVIDED BY ONLINE CONSUMER REVIEWS] Users typically do not address all features and all feature perspectives in each single review. As the analysis of unstructured review texts can be seen as an instrument of open-ended surveys, users are not forced to assess each feature (e.g., in contrast to structured closed-ended surveys, cf. [87]). For instance, 80% of reviews in the restaurant dataset lack either a sentiment for *food quality, service* or *location*, which constitute frequent item aspects and user contexts. That is, users do not necessarily describe all aspects and contexts being potentially relevant for the assigned star rating. Additionally, review texts might even be bound by length restrictions. In our evaluation, such unknown sentiments of aspects and contexts have to be implicitly assumed as neutral sentiments, which puts the achieved explanatory power further into perspective. For instance, the evaluation on the restaurant dataset yields a Nagelkerke pseudo R-squared of 74.5% (compared to 69.8% for the complete dataset) when applied to the 20% of reviews addressing the three features *food quality, service* as well as *location*. Hence, the explanatory power would further increase, when more or all features would be available in review texts instead of only providing a partial view on selected features.

4) [EXPLANATORY POWER FOR DIFFERENT STAR RATING LEVELS] An analysis of the results for different

rating levels (cf. Table 4) shows that the explanatory power differs considerably between rating levels. To be more precise, 1 star and 5 stars ratings are explained considerably well with Nagelkerke pseudo R-squared values up to 82%. This means that very positive or negative reviews can be explained to a high degree, as these reviews often exhibit a very one-sided (clearly positive or clearly negative) line of argumentation. Conversely, explaining 2-4 star ratings is more challenging as these reviews are more nuanced. Comparing our results across domains, the explanatory power regarding the 3 stars level is notably lower for laptops than for movies and restaurants. A sample-based, manual analysis revealed that the structure of neutral reviews (indicated by 3 stars) for search goods (such as laptops) differs from reviews for experience goods, as it seems that these customers have generally informed themselves in detail about the item prior to purchasing a search good. Thus, they only elaborate on facets differing from their expectations in the textual review. This can be illustrated by the exemplary laptop review "*Poor battery! I was so excited to receive my HP, however the battery would not hold a charge for very long. I returned the product.*", which is associated with a 3 star rating, although the user focuses on negative facts. In contrast, almost all neutral reviews of restaurants and movies highlight both positive and negative experiences.

5) [HIGH AMOUNT OF ANALYZED REVIEWS, ITEMS AND USERS] In our evaluation we encompass a high number of users (e.g., approx. 234,000 for the restaurant dataset; cf. Table 2) and items (e.g., approx. 10,000 for the restaurant dataset) per domain (cf. Section 4.1). Additionally, these datasets contain various types of users and items. For instance, the restaurant dataset contains reviews of bars as well as cafes and luxurious restaurants. In contrast, analyses such as surveys or interviews are often limited not only by volume (i.e., the number of users and items being lower), but also in variety (i.e., users and items are of similar types). In comparison to our evaluation, such analyses focus on smaller subsets of similar users or items which could lead to an even higher explanatory power as 'specialized' coefficients explain the star ratings for specific user or item groups better [84]. Conversely, the coefficients and the explanatory power are determined considering all users and items per dataset, ensuring general insights and high validity due to the high number and diversity of reviews, items and users.

Ad RQ2: The results regarding RQ2 show that each feature perspective does indeed contribute to the explanatory power demonstrating the importance of a broader and differentiated view when explaining star ratings. This contributes to our opening question (cf. Section 1), viz., why users rated an item the way they did. Moreover, this finding shows that our research poses a substantial progress compared to existing approaches which address (certain features of) at most two feature perspectives. In particular, we emphasize that the contributions of both object-centered feature perspectives as well as person-centered feature perspectives are remarkable and thus both types of feature perspectives are important to understand star ratings. Consequently, these findings further support the MPAA model [24].

To further substantiate our findings, we tested whether the contribution of each feature perspective is statistically significant. To this end, we compared the unified model (containing all four feature perspectives) with restricted (nested) models (containing only the three feature perspectives) by means of the Bayesian information criterion (BIC) and the likelihood ratio test (LRT) [88,89]. The BIC is particularly suited for the large datasets used in our analyses as it takes into account both the number of independent variables as well as the sample size. Here, the unified model yielded a decrease in the BIC value of at least 978 compared to the restricted model for each feature perspective and all domains, whereby a BIC decrease of 10 indicates 'strong evidence' that the model with lower BIC value is preferred [90]. In that line, the comparisons with the LRT yielded that the unified model is a better model compared to all four restricted models with $p < 10^{-9}$ on all considered domains. Thus, each feature perspective contributes significantly to explaining star ratings, but by analyzing partial Nagelkerke R-squared values in RQ2, we could additionally obtain valuable quantitative results with respect to these feature perspectives as discussed in the following.

The contributions of each feature perspective are comparable across all three domains. Item aspects contribute the most to the explanatory power (e.g., 49.0% for the restaurant domain) since this perspective captures the user's direct perception and the actual experience with an item. For perspectives regarding user and item characteristics, user characteristics contribute more to the explanation of star ratings than item characteristics for the experience goods restaurants and movies

(e.g., 9.8% vs. 1.7% for the restaurant domain). For the search good laptops, item characteristics and user characteristics have an almost equally high contribution in the dataset (8.9% vs. 9.0%). These results can be directly attributed to search goods being more clearly defined by their characteristics (e.g., the capacity of the working memory in the domain of laptops), while experience goods can only be actually assessed after experiencing the item [59]. The particularly low contribution of item characteristics in the restaurant domain might also be due to the fact that users mainly attend the type of restaurants they typically enjoy. For instance, users who dislike Italian food usually will not visit an Italian restaurant. To be more precise, while the item characteristic 'Italian' would be relevant for those users when choosing a restaurant, it rarely comes into effect when writing a review. This indicates that not all features being relevant for purchase decisions are necessarily relevant when explaining star ratings. Additionally, the results show that the feature perspective user contexts has less contribution regarding explanatory power in the domain of laptops compared to movies and restaurants. Since laptops constitute a search good, purchase decisions are arguably more rational and planned in advance. Consequently, it is reasonable that (situational) user contexts do not strongly influence star ratings in this domain.

6 Implications for Research and Practice

Overall, the results and discussions of our evaluation show that (a) the proposed feature perspectives are able to explain star ratings considerably well opening the way for a comprehensive understanding of star ratings and (b) each individual feature perspective contributes to the explanatory power. These findings have implications for both scientific research and practical applications, which are outlined in the following.

6.1 Implications for Research

From a theoretical point of view, our research has the following implications.

1) [SYSTEMATIC AND COMPREHENSIBLE EXPLANATIONS OF STAR RATINGS] Based on the MPAA model, we systematically derived different feature perspectives each consisting of easy-to-interpret features,

which were extracted from review texts. Studying such easy-to-interpret features of different perspectives is a key driver to comprehensibly explaining star ratings which is affirmed by the high overall explanatory power and the substantial contribution of each feature perspective. That is, each feature perspective significantly improves the explanatory power compared to a model restricted to the other three considered feature perspectives. Hence, this work poses a first important step to enable a theoretical foundation – starting from the MPAA model – further research can enhance and use for systematic and comprehensible explanations of star ratings.

2) [ANALYSIS OF DIFFERENT STEPS OF THE RATING SCALE] Moreover, our approach is capable of determining the importance of the features regarding different steps of the rating scale. While prior research has mainly focused on the explanation of 1 star and 5 star ratings [12], our analysis instead also reveals to what degree the features influence star ratings for each rating level within the rating scale. Recent research has initially acknowledged this consideration by separately analyzing reviews with specific star ratings [62]. By aligning to our approach, researchers are now able to examine which features are important for explaining fine-grained differentiations between rating levels (e.g., 4 star and 5 star ratings) in an overall explanatory model.

3) [INCORPORATING CROSS-DOMAIN AND DOMAIN-SPECIFIC INSIGHTS] Our evaluation results show that the different feature perspectives enable to derive cross-domain and domain-specific insights regarding star ratings of online consumer reviews. This can be vital for researchers focusing on cross-domain marketing or domain-independent analysis of user preferences. In particular, our results suggest, that user characteristics such as agreeableness have a positive impact on star ratings in all three analyzed domains while the impact of a user's openness on star ratings varies depending on the particular domain. Hence, these researchers might benefit from incorporating user characteristics in their analyses to explore such relationships and to better understand users. Similarly, research focusing on specific application domains (e.g., hospitality and tourism management) might benefit from domain-specific insights.

4) [UTILIZING MULTIPLE FEATURE PERSPECTIVES FOR OTHER RESEARCH STRANDS] Further, the promising results when utilizing multiple feature perspectives could inspire other research strands

analyzing user assessments. For instance, research in the field of recommender systems mainly use individual feature perspectives to generate personalized item recommendations. That is, content-based recommender systems are largely based on item characteristics [28] while context-aware recommender systems focus on context features [78]. Although predictive and explanatory analysis have different objectives [17], we are confident that our findings encourage researchers to incorporate multiple feature perspectives in recommender systems and other research strands.

5) [USING DIFFERENT FEATURE PERSPECTIVES TO EXPLAIN RECOMMENDATIONS] In addition, works that aim to explain recommendations to the users have gained higher attention in the past years (e.g., [91]). Nevertheless, as popular and widely applied recommender systems infer recommendations based on latent factors (e.g., matrix factorization), it is not straightforward to present meaningful and comprehensible explanations to a user for provided recommendations. Thereby, existing literature tries to explain these recommendations by inferring similarities based on latent factors or by examining item statistics (e.g., movies being popular in a certain region) [92]. With this in mind, our findings could inspire researchers in such fields by analyzing review texts and leveraging easy-to-interpret features and feature perspectives to explain recommended items in a comprehensible manner.

6.2 Implications for Practice

Analyzing star ratings based on different feature perspectives enables consumers, web portals and businesses to leverage the versatile information from online consumer reviews. This allows well-founded and advantageous actions in practical business applications.

1) [GENERATING MEANINGFUL ITEM SUMMARIZATIONS IN WEB PORTALS] By means of our explanatory analysis comprising features of multiple feature perspectives, web portal providers can use the unified model to detect meaningful features and feature perspectives which are important for explaining star ratings (i.e., features with very high or low coefficients). In this line, these meaningful features can be used for summarizing review texts and are particularly relevant for users when forming attitudes about items (e.g., cf. [9]). Consumers might benefit from both individual review

summaries as well as structured summaries encompassing many user reviews of an item. Similarly, and in line with [93], the analysis of different feature perspectives can be used to identify and highlight "informative or representative" review texts, based on the detected meaningful features, which are able to explain star ratings especially well. In that way, users might be more satisfied with the (summarized or selected) information provided by a web portal.

2) [REDUCING USER QUERIES FOR MULTI-CRITERIA RATING SYSTEMS] Furthermore, the derived explanations are valuable for web portal providers that maintain multi-criteria rating systems. Such systems are based on explicit user queries where users are asked to rate specific features after experiencing an item. While a plethora of queries with different features would be possible, answering such queries is an additional effort for users. In order to not discourage users, it is thus only feasible to ask (very) few queries. As the unified model comprises various features of versatile perspectives, the derived explanations enable to identify the most important features (i.e., features with high or low regression coefficients) and perspectives for the users' star ratings. Therefore, web portal providers could focus on important features (e.g., regarding a domain or a group of items) and thereby improve the return of each user query.

3) [IMPROVING ITEMS THROUGH IDENTIFIED USER CRITICISM] In addition, applying the unified model enables businesses to identify features with (very) high and low regression coefficients. This includes features which are often subject to criticism and have a high negative impact on star ratings as well as features which exhibit a positive impact on star ratings. With regard to the increasing volume of user generated content of EWOM and in particular online consumer reviews, aligning to our approach enables businesses to assess these critical features in an automated manner, based on a large review basis and with the possibility to distinguish fine-grained differentiations between rating levels. By analyzing the online consumer reviews regarding these critical features, precise and substantial criticism (e.g., which is expressed in several reviews) can be identified. Consequently and in line with works in the field of EWOM such as [11], businesses are then able to systematically and selectively address the identified criticism, evolve their items, create ideas for new items or new

business models, and prospectively improve the user experience and thus users' item assessments (e.g., star ratings).

4) [USER CHARACTERISTICS ALLOW FOR BETTER CONSUMER UNDERSTANDING] Finally, our results indicate that the feature perspective user characteristics, which has been hardly considered in prior research explaining star ratings, is a key factor in online item assessments and therefore, exhibit high potential for practitioners. Our findings yield strong relations between users' star ratings and the users' personality traits, substantiating and significantly extending the basic findings of [55] on other domains. Therefore, web portal providers, which focus on recommending relevant items to users, as well as businesses providing services or products could benefit from more accurate and comprehensive analyses of consumer behavior by considering the feature perspective user characteristics. For instance, marketing campaigns could target consumers with user characteristics positively influencing star ratings. This could increase the average star rating and thus the revenue of a business [15].

7 Conclusion, Limitations and Future Work

Many web portals such as Amazon or TripAdvisor provide user assessments in form of star ratings and textual reviews. Both research and practice have acknowledged the importance of explaining star ratings. Based upon the existing body of knowledge on this topic, our work is the first to leverage the four feature perspectives item characteristics, item aspects, user characteristics and user contexts in a unified model to explain star ratings in online consumer reviews. We evaluated this model on three large real-world review datasets from the domains of restaurants, laptops and movies using the GOPM. Our results show that these feature perspectives are indeed able to explain star ratings considerably well. Moreover, the evaluation shows that the feature perspectives *item aspects* and *user characteristics* have the highest importance in terms of explanatory power on the star ratings.

Nevertheless, the work at hand has some limitations which could be a starting point for further research. Firstly, the evaluation was conducted on three large datasets from different domains. However, evaluating the unified model on other domains could substantiate and broaden our

findings. Secondly, we operationalized three feature perspectives using a state-of-the-art deep learning language model for analyzing review texts. Nevertheless, other ways of operationalizing feature perspectives (e.g., social identity as user characteristics) could be used to analyze and extend the findings. Thirdly, analyzing interdependent moderator effects between the considered feature perspectives (e.g., do user characteristics influence the effect of item characteristics on star ratings) would be interesting and could complement our findings. Lastly, further analyses of different item or user groups might enable additional insights on the explanation of star ratings and further strengthen the findings of this research.

References

[1] eMarketers, Digital Buyers Worldwide, 2016-2021 (Billions, % Change and % of Internet Users), 2018. https://www.emarketer.com/Chart/Digital-Buyers-Worldwide-2016-2021-billions-change-of-internet-users/215140 (accessed 31 July 2021).
[2] S.M. Mudambi, D. Schuff, What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com, MIS Quarterly 34 (2010) 185. https://doi.org/10.2307/20721420.
[3] X. Li, Revealing or Non-Revealing: The Impact of Review Disclosure Policy on Firm Profitability, MIS Quarterly 41 (2017) 1335–1345. https://doi.org/10.25300/MISQ/2017/41.4.14.
[4] Z. Xiang, Z. Schwartz, J.H. Gerdes, M. Uysal, What Can Big Data and Text Analytics Tell Us about Hotel Guest Experience and Satisfaction?, International Journal of Hospitality Management 44 (2015) 120–130. https://doi.org/10.1016/j.ijhm.2014.10.013.
[5] W. Jabr, B. Liu, D. Yin, H. Zhang, MIS Quarterly Research Curation on Online Word-of-Mouth Research Curation Team (2020).
[6] B. von Helversen, K. Abramczuk, W. Kopeć, R. Nielek, Influence of consumer reviews on online purchasing decisions in older and younger adults, Decision Support Systems 113 (2018) 1–10. https://doi.org/10.1016/j.dss.2018.05.006.
[7] C. Yi, Z. Jiang, X. Li, X. Lu, Leveraging User-Generated Content for Product Promotion: The Effects of Firm-Highlighted Reviews, Information Systems Research 30 (2019) 711–725. https://doi.org/10.1287/isre.2018.0807.
[8] X. Liu, D. Lee, K. Srinivasan, Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning, Journal of Marketing Research 56 (2019) 918–943. https://doi.org/10.1177/0022243719866690.
[9] J. Feng, X. Li, X. Zhang, Online Product Reviews-Triggered Dynamic Pricing: Theory and Evidence, Information Systems Research 30 (2019) 1107–1123. https://doi.org/10.1287/isre.2019.0852.
[10] A.A. Choi, D. Cho, D. Yim, J.Y. Moon, W. Oh, When Seeing Helps Believing: The Interactive Effects of Previews and Reviews on E-Book Purchases, Information Systems Research 30 (2019) 1164–1183. https://doi.org/10.1287/isre.2019.0857.
[11] M. Siering, C. Janze, Information Processing on Online Review Platforms, Journal of Management Information Systems 36 (2019) 1347–1377. https://doi.org/10.1080/07421222.2019.1661094.
[12] W. Jabr, Y. Cheng, K. Zhao, S. Srivastava, What Are They Saying? A Methodology for Extracting Information from Online Reviews, in: Proceedings of the 39th International Conference on Information Systems, 2018.

[13] D. Yin, S.D. Bond, H. Zhang, Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews, MIS Quarterly 38 (2014) 539–560. https://doi.org/10.25300/MISQ/2014/38.2.10.
[14] Chau, Xu, Business Intelligence in Blogs: Understanding Consumer Interactions and Communities, MIS Quarterly 36 (2012) 1189. https://doi.org/10.2307/41703504.
[15] D. Gutt, J. Neumann, S. Zimmermann, D. Kundisch, J. Chen, Design of review systems – A strategic instrument to shape online reviewing behavior and economic outcomes, The Journal of Strategic Information Systems 28 (2019) 104–117. https://doi.org/10.1016/j.jsis.2019.01.004.
[16] S. Gensler, F. Völckner, M. Egger, K. Fischbach, D. Schoder, Listen to Your Customers: Insights into Brand Image Using Online Consumer-Generated Product Reviews, International Journal of Electronic Commerce 20 (2015) 112–141. https://doi.org/10.1080/10864415.2016.1061792.
[17] G. Shmueli, To Explain or to Predict?, Statistical Science 25 (2010) 289–310.
[18] R.P. Karumur, T.T. Nguyen, J.A. Konstan, Exploring the Value of Personality in Predicting Rating Behaviors: A Study of Category Preferences on Movielens, in: Proceedings of the Tenth ACM Conference on Recommender Systems - RecSys '16, Boston Massachusetts USA, ACM Press, New York, NY, USA, 2016, pp. 139–142.
[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
[20] M. Binder, B. Heinrich, M. Klier, A. Obermeier, A. Schiller, Explaining the Stars: Aspect-based Sentiment Analysis of Online Customer Reviews, in: Proceedings of the 27th European Conference on Information Systems, 2019.
[21] M.M. Tunc, H. Cavusoglu, S. Raghunathan, Online Product Reviews: Is a Finer-Grained Rating Scheme Superior to a Coarser One?, MIS Quarterly 45 (2021) 2193–2234. https://doi.org/10.25300/MISQ/2022/15586.
[22] S. Chatterjee, Explaining Customer Ratings and Recommendations by Combining Qualitative and Quantitative User Generated Contents, Decision Support Systems (2019) 14–22.
[23] M. Siering, A.V. Deokar, C. Janze, Disentangling Consumer Recommendations: Explaining and Predicting Airline Recommendations Based on Online Reviews, Decision Support Systems 107 (2018) 52–63.
[24] J.B. Cohen, A. Reed, A Multiple Pathway Anchoring and Adjustment (MPAA) Model of Attitude Generation and Recruitment, Journal of Consumer Research 33 (2006) 1–15.
[25] L.R. Glasman, D. Albarracín, Forming attitudes that predict future behavior: a meta-analysis of the attitude-behavior relation, Psychol. Bull. 132 (2006) 778–822. https://doi.org/10.1037/0033-2909.132.5.778.
[26] I. Ajzen, M. Fishbein, The Influence of Attitudes on Behavior, in: B.T. Johnson, D. Albarracin, M.P. Zanna (Eds.), The handbook of attitudes, Lawrence Erlbaum Associates, Mahwah, N.J, 2005, pp. 173–221.
[27] T.D. Wilson, S. Lindsey, T.Y. Schooler, A model of dual attitudes, Psychol. Rev. 107 (2000) 101–126. https://doi.org/10.1037/0033-295x.107.1.101.
[28] F. Ricci, L. Rokach, B. Shapira, Recommender Systems: Introduction and Challenges, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, Boston, MA, 2015, pp. 1–34.
[29] T. Radojevic, N. Stanisic, N. Stanic, Inside the Rating Scores: A Multilevel Analysis of the Factors Influencing Customer Satisfaction in the Hotel Industry, Cornell Hospitality Quarterly 58 (2017) 134–164. https://doi.org/10.1177/1938965516686114.
[30] A.K. Jha, S. Shah, Social Influence on Future Review Sentiments: An Appraisal-Theoretic View, Journal of Management Information Systems 36 (2019) 610–638. https://doi.org/10.1080/07421222.2019.1599501.
[31] K.C. Briggs, Myers-Briggs Type Indicator, Consulting Psychologists Press Palo Alto, CA, 1976.

[32] L.R. Goldberg, An Alternative "Description of Personality": The Big-Five Factor Structure, Journal of Personality and Social Psychology 59 (1990) 1216–1229.
[33] C.K. Manner, W.C. Lane, Who posts online customer reviews? The role of sociodemographics and personality traits, Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior 30 (2017) 1–24.
[34] M. Husnain, I. Qureshi, T. Fatima, W. Akhtar, The impact of electronic word-of-mouth on online impulse buying behavior: The moderating role of Big 5 personality traits, Journal of Accounting & Marketing 5 (2016) 190–209.
[35] R. Hu, P. Pu, Exploring Relations between Personality and User Rating Behaviors, in: UMAP Workshops, 2013.
[36] P. Adamopoulos, A. Ghose, V. Todri, The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms, Information Systems Research 29 (2018) 612–640. https://doi.org/10.1287/isre.2017.0768.
[37] P.G. Campos, N. Rodriguez-Artigot, I. Cantador, Extracting Context Data from User Reviews for Recommendation: A Linked Data Approach, in: ComplexRec@ RecSys, 2017, pp. 14–18.
[38] Q.B. Liu, E. Karahanna, The Dark Side of Reviews: The Swaying Effects of Online Product Reviews on Attribute Preference Construction, MISQ 41 (2017) 427–448. https://doi.org/10.25300/MISQ/2017/41.2.05.
[39] U. Panniello, M. Gorgoglione, A. Tuzhilin, In CARS We Trust: How Context-Aware Recommendations Affect Customers' Trust And Other Business Performance Measures Of Recommender Systems, in: CARS, 2015.
[40] P. Potash, A. Rumshisky, Recommender System Incorporating User Personality Profile through Analysis of Written Reviews, in: EMPIRE@ RecSys, 2016, pp. 60–66.
[41] J. Qiu, C. Liu, Y. Li, Z. Lin, Leveraging Sentiment Analysis at the Aspects Level to Predict Ratings of Reviews, Information Sciences 451 (2018) 295–309.
[42] N.R. Draper, H. Smith, Applied Regression Analysis, Third edition, Wiley, New York, Chichester, Weinheim, Brisbane, Singapore, Toronto, 1998.
[43] Y. Levy, T. J. Ellis, A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research, Informing Science: The International Journal of an Emerging Transdiscipline 9 (2006) 181–212. https://doi.org/10.28945/479.
[44] P. Chen, Y. Ge, Y. Hong, Y. Liu, The Impact of Rating System Design on Opinion Sharing, in: Proceedings of the 38th International Conference on Information Systems, 2017.
[45] J. Linshi, Personalizing Yelp Star Ratings: A Semantic Topic Modeling Approach, Yale University (2014).
[46] E. Lacic, D. Kowald, E. Lex, High Enough?: Explaining and Predicting Traveler Satisfaction Using Airline Reviews, in: Proceedings of the 27th ACM Conference on Hypertext and Social Media, 2016, pp. 249–254.
[47] Y. Guo, S.J. Barnes, Q. Jia, Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation, Tourism Management 59 (2016) 467–483. https://doi.org/10.1016/j.tourman.2016.09.009.
[48] Y. Liu, T. Teichert, M. Rossi, H. Li, F. Hu, Big Data for Big Insights: Investigating Language-specific Drivers of Hotel Satisfaction with 412,784 User-generated Reviews, Tourism Management 59 (2017) 554–563. https://doi.org/10.1016/j.tourman.2016.08.012.
[49] Y. Luo, R.L. Tang, Understanding Hidden Dimensions in Textual Reviews on Airbnb: An Application of Modified Latent Aspect Rating Analysis (LARA), International Journal of Hospitality Management 80 (2019) 144–154.
[50] Q. Gan, B.H. Ferns, Y. Yu, L. Jin, A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews, Journal of Quality Assurance in Hospitality & Tourism 18 (2017) 465–492.
[51] Q. Gan, Y. Yu, Restaurant Rating: Industrial Standard and Word-of-Mouth -- A Text Mining and Multi-dimensional Sentiment Analysis, in: 48th Hawaii International Conference on System Sciences (HICSS), 2015, pp. 1332–1340.

[52] Q. Ye, H. Li, Z. Wang, R. Law, The Influence of Hotel Price on Perceived Service Quality and Value in E-Tourism, Journal of Hospitality & Tourism Research 38 (2014) 23–39. https://doi.org/10.1177/1096348012442540.
[53] X. Xu, Does Traveler Satisfaction Differ in Various Travel Group Compositions? Evidence from Online Reviews, International Journal of Contemporary Hospitality Management 30 (2018) 1663–1685.
[54] S. Chatterjee, P. Mandal, Traveler preferences from online reviews: Role of travel goals, class and culture, Tourism Management 80 (2020) 104108.
[55] M. Han, Examining the effect of reviewer expertise and personality on reviewer satisfaction: An empirical study of TripAdvisor, Computers in Human behavior (2020) 106567.
[56] B.B. McShane, D. Gal, Statistical significance and the dichotomization of evidence, Journal of the American Statistical Association 112 (2017) 885–895.
[57] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management 35 (2015) 137–144.
[58] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 Task 4: Aspect Based Sentiment Analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35.
[59] M. Schmalz, M. Carter, J.H. Lee, It's Not You, It's Me: Identity, Self–Verification, and Amazon Reviews, The DATA BASE for Advances in Information Systems 49 (2018).
[60] J. Ni, J. Li, J. McAuley, Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 188–197.
[61] S. Debortoli, O. Müller, I. Junglas, J. Vom Brocke, Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial, CAIS 39 (2016) 110–135. https://doi.org/10.17705/1CAIS.03907.
[62] D. Keller, M. Kostromitina, Characterizing non-chain restaurants' Yelp star-ratings: Generalizable findings from a representative sample of Yelp reviews, International Journal of Hospitality Management 86 (2020) 102440.
[63] G. Heinze, C. Wallisch, D. Dunkler, Variable selection-a review and recommendations for the practicing statistician, Biometrical Journal 60 (2018) 431–449.
[64] D. Ramage, C.D. Manning, S. Dumais, Partially labeled topic models for interpretable text mining, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, 2011.
[65] V. Vallurupalli, I. Bose, Exploring thematic composition of online reviews: A topic modeling approach, Electron Markets 30 (2020) 791–804. https://doi.org/10.1007/s12525-020-00397-5.
[66] I. Tenney, D. Das, E. Pavlick, BERT Rediscovers the Classical NLP Pipeline, Association for Computational Linguistics, 2019.
[67] H. Xu, B. Liu, L. Shu, P. Yu, BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2324–2335.
[68] S.-T. Park, W. Chu, Pairwise Preference Regression for Cold-Start Recommendation, in: Proceedings of the Third ACM Conference on Recommender Systems, 2009, pp. 21–28.
[69] J.W. Pennebaker, L.A. King, Linguistic styles: Language use as an individual difference, Journal of Personality and Social Psychology 77 (1999) 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296.
[70] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artificial Intelligence Review 53 (2020) 2313–2339.

[71] G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D.J. Stillwell, L.H. Ungar, M.E.P. Seligman, Automatic personality assessment through social media language, Journal of Personality and Social Psychology 108 (2015) 934.
[72] R.M. O'brien, A Caution Regarding Rules of Thumb for Variance Inflation Factors, Quality & Quantity 41 (2007) 673–690. https://doi.org/10.1007/s11135-006-9018-6.
[73] L. Yu, J. Huang, G. Zhou, C. Liu, Z.-K. Zhang, TIIREC: A tensor approach for tag-driven item recommendation with sparse user generated content, Information Sciences 411 (2017) 122–135. https://doi.org/10.1016/j.ins.2017.05.025.
[74] M. de Gemmis, P. Lops, C. Musto, F. Narducci, G. Semeraro, Semantics-Aware Content-Based Recommender Systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, Boston, MA, 2015, pp. 119–159.
[75] T.T. Thet, J.-C. Na, C.S. Khoo, Aspect-based sentiment analysis of movie reviews on discussion boards, Journal of Information Science 36 (2010) 823–848. https://doi.org/10.1177/0165551510388123.
[76] J. Wang, J. Li, S. Li, Y. Kang, M. Zhang, L. Si, G. Zhou, Aspect Sentiment Classification with Both Word-Level and Clause-Level Attention Networks, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press, 2018, pp. 4439–4445.
[77] F. Celli, Unsupervised Personality Recognition for Social Network Sites, in: Proceedings of the Sixth International Conference on Digital Society, 2012.
[78] G. Adomavicius, A. Tuzhilin, Context-Aware Recommender Systems, in: F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (Eds.), Recommender Systems Handbook, Springer US, Boston, MA, 2011, pp. 217–253.
[79] N.J.D. Nagelkerke, A Note on a General Definition of the Coefficient of Determination, Biometrika 78 (1991) 691–692.
[80] R.D. McKelvey, W. Zavoina, A Statistical Model for the Analysis of Ordinal Level Dependent Variables, The Journal of Mathematical Sociology 4 (1975) 103–120. https://doi.org/10.1080/0022250X.1975.9989847.
[81] G.S. Maddala, Limited-dependent and Qualitative Variables in Econometrics, Cambridge University Press, 1983.
[82] R. Anderson-Sprecher, Model Comparisons and R 2, The American Statistician 48 (1994) 113–117.
[83] P. Legendre, L.F.J. Legendre, Numerical Ecology, thirdrd Edition, Elsevier, 2012.
[84] H.-J. Kim, M.P. Fay, B. Yu, M.J. Barrett, E.J. Feuer, Comparability of segmented line regression models, Biometrics 60 (2004) 1005–1014.
[85] B. Anastasiei, N. Dospinescu, A model of the relationships between the Big Five personality traits and the motivations to deliver word-of-mouth online, Psihologija 51 (2018) 215–227. https://doi.org/10.2298/psi161114006a.
[86] T. Formánek, R. Tahal, Brand importance across product categories in the Czech Republic, Management & Marketing. Challenges for the Knowledge Society 11 (2016) 341–354. https://doi.org/10.1515/mmcks-2016-0001.
[87] E. Singer, M.P. Couper, Some Methodological Uses of Responses to Open Questions and Other Verbatim Comments in Quantitative Surveys, Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (2017) 115–134. https://doi.org/10.12758/mda.2017.01.
[88] G. Schwarz, Estimating the Dimension of a Model, The Annals of Statistics 6 (1978) 461–464.
[89] Q.H. Vuong, Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses, Econometrica 57 (1989).
[90] A.E. Raftery, Bayesian model selection in social research, Sociological methodology (1995) 111–163.
[91] I. Nunes, D. Jannach, A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems, User Modeling and User-Adapted Interaction 27 (2017) 393–444. https://doi.org/10.1007/s11257-017-9195-0.

[92] N. Tintarev, J. Masthoff, Explaining Recommendations: Design and Evaluation, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, Boston, MA, 2015, pp. 353–382.
[93] J. McAuley, J. Leskovec, Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text, in: Proceedings of the 7th ACM Conference on Recommender Systems, 2013, pp. 165–172.

3 Enhanced Approaches in RS and ABSA

"We must enhance the light, not fight the darkness."
Aaron David Gordon (*1856; †1922)

3.1 Paper: Something's Missing? A Procedure for Extending Item Content Data Sets in the Context of Recommender Systems

Current Status	**Citation**
This paper is **accepted** and **published** in Volume 24, Issue 1 in the journal *Information Systems Frontiers*.	Heinrich, B., M. Hopf, D. Lohninger, A. Schiller and M. Szubartowicz (2022b). "Something's Missing? A Procedure for Extending Item Content Data Sets in the Context of Recommender Systems" Information Systems Frontiers 24 (1), 267–286.

• Something's Missing? A Procedure for Extending Item Content Data Sets in the Context of Recommender Systems

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

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

Daniel Lohninger, University of Regensburg, Regensburg, Germany, daniel.lohninger@ur.de

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

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

Abstract

The rapid development of e-commerce has led to a swiftly increasing number of competing providers in electronic markets, which maintain their own, individual data describing the offered items. Recommender systems are popular and powerful tools relying on this data to guide users to their individually best item choice. Literature suggests that data quality of item content data has substantial influence on recommendation quality. Thereby, the dimension completeness is expected to be particularly important. Herein resides a considerable chance to improve recommendation quality by increasing completeness via extending an item content data set with an additional data set of the same domain. This paper therefore proposes a procedure for such a systematic data extension and analyzes effects of the procedure regarding items, content and users based on real-world data sets from four leading web portals. The evaluation results suggest that the proposed procedure is indeed effective in enabling improved recommendation quality.

1 Introduction

In line with the emergence and proliferation of the internet, e-commerce has developed into a major disruptor for retail business. Indeed, in 2020, retail e-commerce sales worldwide are estimated to hit $4.2 trillion, with its share of global retail reaching 16.1% and rising further to 22% in 2023 (Statista 2019). This rapid development of e-commerce has implied a swiftly increasing number of competing providers in electronic markets (e.g., *Amazon* and *Walmart* in general retail, *Booking.com* and *HRS* in hotel bookings, *Yelp* and *TripAdvisor* in restaurant bookings). Providers – even of the same domain – maintain their own, individual data sets containing information regarding the offered items (e.g., products or services), which usually vary in their attributes (content) to describe even the same items. For instance, *Booking.com* provides detailed data on location score and furniture of hotels, which is not offered by *HRS*. This data as well as the recommender systems commonly present on such e-commerce platforms aim at guiding users to their individually best item choice, improving user stickiness and increasing platform revenue (Zhou 2020). Such supporting systems are mandatory as customers regularly need to make a choice between a plethora of items (e.g., songs, movies, restaurants, hotels) on e-commerce platforms (Kamis et al. 2010; Levi et al. 2012; Richthammer and Pernul 2018; Tang et al. 2017; Vargas-Govea et al. 2011). It is thus hardly surprising that recommender systems in particular have been established as one of the most powerful and popular tools in the field of e-commerce in recent years (Ricci et al. 2015a; Scholz et al. 2017; Smith and Linden 2017).

As recommender systems are data-driven tools, the quality of the data which a recommender system is based on is assessed to be one of the issues recommender systems research is strongly involved with (Bunnell et al. 2019) and may have substantial influence on the resulting recommendations (Picault et al. 2011; Sar Shalom et al. 2015). Here, data quality is a multidimensional construct comprising several dimensions such as accuracy, completeness and currency of data (Batini and Scannapieco 2016; Pipino et al. 2002; Wand and Wang 1996), with each dimension providing a distinct view on data quality (e.g., Heinrich et al. 2018). For recommender systems examining the item content data (attributes and attribute values of items), achieving a more complete view on these items seems to be especially important (Adomavicius and Tuzhilin 2005; Picault et al. 2011), as "some representations capture only certain aspects of the content, but there are many others that would influence a user's experience" (Picault et al. 2011). This means that the data quality dimension completeness is of particular relevance for recommender systems.

Herein resides a considerable chance to improve recommendation quality by increasing completeness via extending an item content data set (e.g., from an e-commerce platform such as *TripAdvisor*) with additional attributes and

attribute values from another data set in the same domain (e.g., from an e-commerce platform such as *Yelp*). This opportunity is particularly promising for search portals offering a meta view by compiling information from various platforms (e.g., *trivago.com*), which currently simply juxtapose the data and do not use an extended data set for the application of a recommender system. Yet, how to systematically achieve more complete item content data sets and realize the expected advantages for recommender systems is left unanswered in existing research. Thus, the paper at hand investigates the following research question:

How can an item content data set be systematically extended with respect to the data quality dimension completeness, aiming to improve recommendation quality?

As recommender systems are an important category of decision support systems (Power et al. 2015), this research is in line with recent works which have revealed a significant impact of data quality dimensions, especially completeness, on data-driven decision support systems (e.g., Feldman et al. 2018; Heinrich et al. 2019; Woodall et al. 2015).
The remainder of the paper is organized as follows. In the next section, the general and theoretical background as well as the related work are discussed. Thereafter, a procedure for the systematic extension of an item content data set with attributes and attribute values from another item content data set is presented, providing the basis for determining recommendations. In the fourth section, the proposed procedure is evaluated in two e-commerce real-world scenarios and resulting effects on recommendation quality are analyzed. The final section summarizes the work and discusses limitations as well as directions for future research.

2 Foundation

This section first discusses the positioning of recommender systems in the field of decision support systems in e-commerce as general background of our research. The second part of this section presents a theoretical model regarding the relationship between data quality and decision support systems – especially recommender systems – based on a discussion of existing literature. The third part of the section discusses related work and identifies the research gap addressed by this paper.

2.1 General Background

Recommender systems have become a highly relevant category of decision support systems (Power et al. 2015). In particular, in e-commerce, recommender systems are often necessary as users regularly need to make decisions for purchase, consumption or utilization of items (e.g., songs, movies, restaurants or hotels) from a plethora of possible alternatives available in information systems (IS) on e-commerce platforms (Kamis et al. 2010; Levi et al. 2012; Richthammer and Pernul 2018; Tang et al. 2017; Vargas-Govea et al. 2011).
More precisely, the high number of items together with the high number of users on e-commerce platforms lead to the problem of information overload, which is widely discussed by many researchers in the past decades and thus, constitutes a major subject of IS research in fields such as e-commerce (Lu et al. 2015) or management of business organizations (Edmunds and Morris 2000). In particular, information overload denotes the phenomenon regarding an individual's ability to appropriately cope with solving problems (e.g., making a choice) when more information is available than the individual can assimilate (Edmunds and Morris 2000). This is, users often do not have the skills and experience to adequately evaluate the large number of available alternatives for making their choice (Ricci et al. 2015b; Scholz et al. 2017). The resulting problem leaves users of e-commerce IS unable to make effective decisions due to this large volume of information (e.g., items) to which users are exposed to (Hasan et al. 2018; Lu et al. 2015; Richthammer and Pernul 2018; Scholz et al. 2017). In order to address the problem of information overload, the literature suggests for IS providers in e-commerce to incorporate decision support systems, in particular recommender systems, to assist users in their decision-making (Bunnell et al. 2019; Karimova 2016; Lu et al. 2015). Therefore, recommender systems aim at individually preselecting smaller sets of relevant items for each single user (i.e., information filtering; cf. Lu et al. 2015) to allow for good decision-making in a personalized and comfortable way avoiding to overwhelm the user (Manca et al. 2018).
Here, recommender systems are especially suitable to tackle the information overload problem, since they constitute data-driven systems, which enables them to individually support each user's decision-making in an automated manner (Bunnell et al. 2019; Karumur et al. 2018; Lu et al. 2015). A variety of IS research aims to tackle the information overload problem in the field of e-commerce by developing different approaches for recommender systems (e.g., Content-Based Filtering; cf. Aggarwal 2016; Jannach et al. 2012; Ricci et al. 2015a). In particular, recommender systems process different types of data (e.g., user rating data or item content data) in order to derive the individual users' preferences, which are stored in a user profile, based on data such as users' historical evaluations of other items (cf. Peska and Vojtas 2015; Ricci et al. 2015a). To enable recommendations of high precision, the matching of the

user profile against item profiles (i.e., the content data of an item) or against other user profiles is highly relevant (Ricci et al. 2015a). This further emphasizes the key role of data (e.g., item content data) for recommender systems to enable individualized decision support for a large number of users in e-commerce settings (e.g., during shopping experiences on e-commerce websites; cf. Heinrich et al. 2019; Kamis et al. 2010).
In e-commerce, recommender systems not only assist users and make their experience on e-commerce platforms more comfortable, but they also create business value for the IS providers (Bunnell et al. 2019). By integrating recommender systems into a wide variety of e-commerce activities such as browsing, purchasing, rating or reviewing items, the resulting diversity of generated data (e.g., item content data, user rating data or click-stream data) can be used for modeling of user profiles and thus support certain marketing activities such as cross-selling, advertising or product promotion (Karimova 2016; Lu et al. 2015). It is thus hardly surprising that in recent years, recommender systems as data-driven tools have emerged to be among the most frequently applied decision support systems in the field of IS in e-commerce (Ricci et al. 2015a; Scholz et al. 2017; Smith and Linden 2017).
As recommender systems support user choices mainly on the basis of data, it seems promising to investigate how the data quality (e.g., completeness of item content data) influences the quality of recommender systems in the field of e-commerce.

2.2 Theoretical Background

The systematic procedure presented in this paper aims to contribute to further research investigating the relationship between data quality and (data-driven) decision support systems. At first glance, it might seem natural and obvious to suggest that more data always has a positive influence on decision support (especially when provided by a system). However, research in different areas shows that more data does not always lead to better results of decision support systems in general (e.g., when selecting features based on which a decision is obtained; cf. Mladenić and Grobelnik 2003; Vanaja and Mukherjee 2019), as different data sets (e.g., with more or fewer attributes) may lead to varying results of decision support. Thus, the impact of the data quality of data values on different evaluation criteria of decision support systems such as decision quality or data mining outcome has been studied in existing literature (e.g., Bharati and Chaudhury 2004; Blake and Mangiameli 2011; Feldman et al. 2018; Ge 2009; Heinrich et al. 2019; Woodall et al. 2015). Yet, this research neither focuses on how to systematically achieve more complete item content data sets nor on how to define a well-founded procedure, but instead tries to explain the relationship between data quality and evaluation criteria of decision support systems. In this regard, such explanatory models are the theoretical background in data quality research which we aim to support by our work. Thus, this background is briefly discussed in the following.
Bharati and Chaudhury (2004) assess the effects of the data quality dimensions accuracy, completeness and currency on the ability of an online analytical processing system to sustain decision-making. Ge (2009) discusses accuracy, completeness and consistency and their impact on decision quality. Blake and Mangiameli (2011) assess the impact of accuracy, completeness, consistency and currency on data mining results in order to support decision-making in companies. Woodall et al. (2015) analyze the impact of completeness on classification outcomes used for supporting users in their decision process. Feldman et al. (2018) propose an analytical framework to investigate the effects of incomplete data sets on a binary classifier that serves for decision support. Heinrich et al. (2019) examine the impact of the amount of available attributes and attribute values on the prediction accuracy of recommender systems.
Summing up, the focus of these papers is to investigate in which way and to what extent improving the quality of data values, especially the dimension completeness, leads to an improvement in evaluation criteria of particular decision support systems. A relevant and widely used category of decision support systems which assists users facing decision-making problems are recommender systems (Porcel and Herrera-Viedma 2010; Power et al. 2015). Based on this and in line with Heinrich et al. (2019), we refer to the theoretical model for describing the relationship between data quality and decision support systems, presented in Fig. 1.

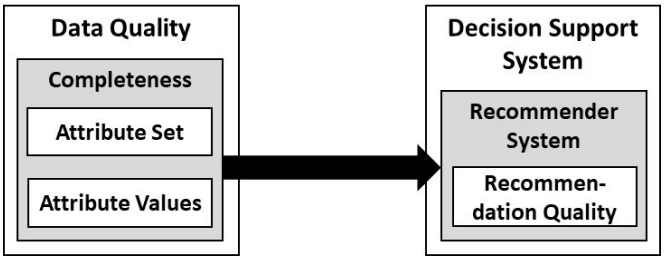

Fig. 1 Theoretical Model (according to Heinrich et al. 2019)

The theoretical model in Fig. 1 indicates a direct relationship between data quality and decision support systems. In particular, the theoretical model suggests this relationship between completeness of item content data (attributes and attribute values) and recommendation quality of recommender systems. With this model as theoretical background, the procedure presented in this paper proposes how to systematically extend items in an item content data set with attributes and attribute values of the same items from a second item content data set in order to gain a more complete view on the considered real-world entities (e.g., movies or restaurants). Thus, this systematic procedure forms the basis for an even more precise and well-founded investigation of the impact of completeness on the recommendation quality of data-driven decision support systems (especially recommender systems) in the future.[1] In particular, it enables theoretical relationships (i.e., similar to Fig. 1) for different data sets to be analyzed in a transparent and comprehensible manner. Furthermore, this procedure can serve as an already evaluated template for future procedures in order to support the investigation of further data quality dimensions (e.g., consistency) in other data-driven decision support systems.

2.3 Related Work and Research Gap

In this section, we present approaches dealing with data extension in the context of recommender systems and analyze relevant works discussing data quality aspects related to recommender systems.[2] Thereafter, we summarize existing contributions and identify the research gap addressed by this paper.

To prepare the related work, we followed the guidelines of standard approaches (e.g., Levy and Ellis 2006). In particular, we performed a literature search on the databases *ACM Digital Library*, *AIS Electronic Library*, *IEEE Xplore*, *ScienceDirect* and *Springer* as well as the proceedings of the *European and International Conference on Information Systems*, the *International Conference on Information Quality* and the *ACM Conference on Recommender Systems*. Subsequently, we examined whether these works represent relevant approaches for our research by reading title, keywords, abstract and summary and also conducted a forward and backward search in order to find further relevant works. After analyzing the resulting papers in detail, eighteen articles were deemed relevant. These papers could be organized within two separate categories, with each category containing nine works.

(1): The first category of works copes with some kind of data extension in the context of recommender systems. For these works, the effect on decision quality and in particular recommendation quality is vital ("fitness for use"). This is a crucial difference to general approaches for data extension (e.g., in the context of data warehouses), where the effect on decision quality is often unclear or difficult to assess. Although all papers of the first category consider data extension and its effect on recommendation quality, none of the approaches describes the systematic extension of an item content data set with additional data from the same domain in the form of a procedure in the context of recommender systems, which is the contribution of our research. Moreover, the approaches differ in the kind of extended data (1A), the entities extended with data (1B) and in the usage of different methods for data extension (1C).

(1A): Several recent articles focus on the extension of data with data from a distinct area, for example, data from different domains such as music and film (cross-domain data sets; Abel et al. 2013; Ntoutsi and Stefanidis 2016; Ozsoy et al. 2016), context information such as time and location (multi-dimensional data sets; Abel et al. 2013; Kayaalp et al. 2009) or data from different social and semantic web sources such as *Wikipedia*, *Facebook* and *Twitter* (heterogeneous data sets; Abel et al. 2013; Bostandjiev et al. 2012; Chang et al. 2018; Kayaalp et al. 2009; Ozsoy et al. 2016). These approaches examine whether the diversity of data types leads to improved recommendation quality but do not systematically extend item content data with additional data from the same domain.

(1B): Other works in literature analyze user profiles from different social networks (Abel et al. 2013; Li et al. 2018; Ozsoy et al. 2016; Raad et al. 2010). The matching user profiles are merged across different networks to produce a positive effect on recommendation quality. However, these works do not focus on item content data at all.

(1C): Finally, some recent works focus on the extension of item or user data from multiple data sources in the context of recommender systems (Abel et al. 2013; Bostandjiev et al. 2012; Bouadjenek et al. 2018; Ozsoy et al. 2016). These approaches rely on tools such as *BlogCatalog*, *Google Social Graph API*, *Google Search API* or *OpenID*, which provide information for the matching of users or items. However, these works do not focus on describing the systematic extension of an item content data set and instead use external, non-transparent methods for data extension, which severely limits their applicability in other scenarios.

---

[1] In this regard, an implementation of the procedure is available on GitHub (GitHub 2020).

[2] Some approaches for data extension with regard to completeness (e.g., cf. Naumann et al. 2004; Bleiholder and Naumann 2008; Scannapieco and Batini 2004) mainly deal with technical issues (e.g., wrapper architecture, database architecture) or model-oriented aspects (e.g., schema mapping, operators, join approaches), which are not within the scope of this work.

(2): The second category of works explicitly recognizes the importance of data quality for recommender systems (Amatriain et al. 2009; Basaran et al. 2017; Berkovsky et al. 2012; Burke and Ramezani 2011; Heinrich et al. 2019). In particular, Heinrich et al. (2019) examine the impact of the number of available attributes and attribute values on prediction accuracy of recommender systems by testing hypotheses but do not provide a procedure for extending an item content data set with additional attributes and attribute values. Further approaches give rise to concepts that deal with data quality issues in the context of recommender systems. For instance, data sources used by a recommender system can be chosen user-dependently as data sparsity and inaccuracy have been identified to impact recommendation quality (Lathia et al. 2009). Sar Shalom et al. (2015) tackle sparsity and redundancy issues by deleting or omitting certain users or items while Pessemier et al. (2010) analyze consumption data such as ratings in regard to currency. Further, Levi et al. (2012) use text mining on user reviews from various sources to alleviate the cold start problem of new users by assigning them to so called context groups.

In summary, none of these works provides a systematic procedure for the extension of a data set with item content data of another data set from the same domain. The works in category (1A) focus on the extension with data from a different area, but they do not target on data representing the *same* items, which is a decisive characteristic of our research. The works in category (1B) do not focus on item content data but instead analyze user profiles from various social networks. In contrast to this, we provide a procedure for the matching and extension of *item* content data. The works in category (1C) use existing tools for data extension, especially for user data. Such an extension is non-transparent, highly dependent on these tools as well as the application scenario and does not allow to support the analysis of theoretical relationships (cf. Fig. 1) between different data sets in a verifiable and comprehensible manner. Additionally, no explicit procedure for extending an item content data set with additional attributes and attribute values in detail is given. The works of the second category analyze the impact of data quality on recommender systems. However, only Heinrich et al. (2019) analyze effects of a more complete view on items by data set extension. Yet, this work does not aim to provide a procedure for the extension of item content data in the context of recommender systems. In contrast, the authors present an explanatory analysis based on hypotheses testing. To conclude, none of these approaches presents a systematic procedure for the extension of a data set with item content data of another data set from the same domain.

3 A Procedure for Extending an Item Content Data Set

In this section, we propose a procedure for the systematic extension of a data set in the context of recommender systems, aiming to improve the quality of the resulting recommendations. We discuss and substantiate in detail how to extend a data set DS1 containing items and item attributes from a certain domain (e.g., movies, restaurants or hotels) by using a data set DS2 containing items and item attributes from the same domain.[3] In particular, items in DS1 are extended with attributes and attribute values of the same items from DS2. This means that in a first step *duplicates have to be detected* before in a second step, the *data sets can be actually integrated into one data set*.

The exact elaboration of these two steps in the context of recommender systems addresses our research question and thus represents the contribution of this paper. In a subsequent step, the resulting data set extension can be evaluated by *determining recommendations* based on the extended data set and assessing the resulting recommendation quality. Since different existing content-based or hybrid recommender systems can be used for this step, it is not a core element of the procedure. The procedure is illustrated in Fig. 2 and described in the following.

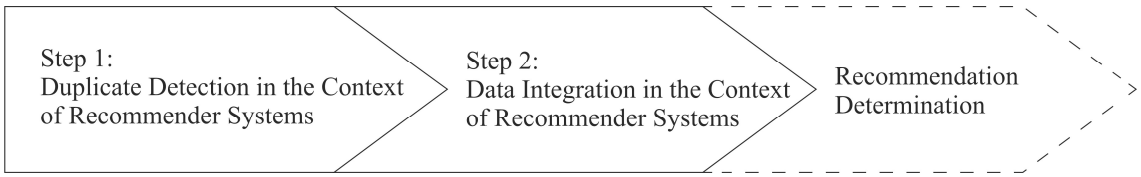

Fig. 2 Procedure to Extend an Item Content Data Set in the Context of Recommender Systems

3.1 Duplicate Detection in the Context of Recommender Systems

An item in a data set DS1 usually has different attributes and attribute values compared to its corresponding duplicate item in a data set DS2 (e.g., because the portals have heterogeneous data policies), making duplicate detection in the context of recommender systems a non-trivial task. Here, duplicate detection is a binary classification of item pairs

[3] If more than two data sets are available, the procedure can be applied iteratively.

(one item from DS1 and one item from DS2) with the two classes *duplicate* and *non-duplicate*. Due to a potentially large number of items per data set, duplicate detection should be carried out in a widely automated manner. To assist this task, literature proposes *similarity measure functions* (SMFs; e.g., the Jaro-Winkler function; Winkler 1990) to determine the similarity of *key attributes* (e.g., "Name" and "Geolocation" of a restaurant) between items from DS1 and DS2, with high similarity values indicating possible duplicates. We propose the following four Tasks 1.1-1.4 to configure and perform duplicate detection, acknowledging peculiarities in the context of recommender systems (cf. Fig. 3).

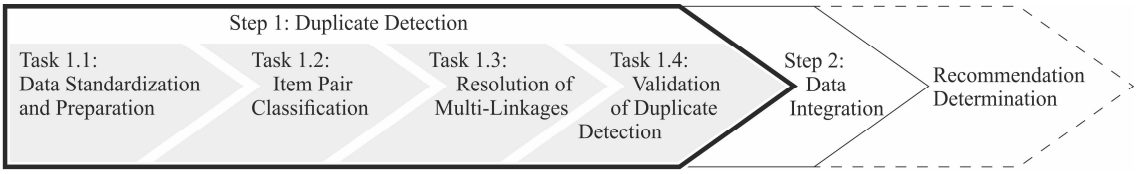

Fig. 3 The Step Duplicate Detection in Detail

In **Task 1.1**, the data for duplicate detection is standardized and prepared. This is necessary because different portals often specify varying values for (key) attributes (e.g., due to heterogeneous data policies). Furthermore, as the data is usually decentrally generated by many different users, these users often enter attribute values on their very own interpretation, leading to data quality problems in e-commerce platforms. These issues make duplicate detection for recommender systems data sets highly complex. For example, one and the same US phone number could be entered as "+1-212-283-1100" in one data set and as "(212) 283-1100" in the other data set. Here, it is clear that a standardization of both phone numbers to "area code: 212, phone number: 2831100" helps determining that these numbers refer to the same phone connection in an automated manner. The standardization of the key attributes can be conducted by utilizing specific parsing tools which standardize the values of the key attributes (e.g., the python package "phonenumbers" for the key attribute "Phone"). After standardization, the values for all key attributes of both data sets DS1 and DS2 are stored in a common standard format. Nevertheless, even after standardization, duplicate items in DS1 and DS2 may differ in key value attributes caused by varying entered values (e.g., "283-100" instead of "283-1100"). Subsequent to standardization, item pairs are prepared for binary classification in the next task. Here, each item from DS1 in combination with each item from DS2 is considered as an item pair. It is clear that most of these pairs are non-duplicates. Therefore, it is beneficial to discard the item pairs which are obvious non-duplicates (e.g., restaurants with a GPS distance larger than 1,000 meters), which is referred to as blocking in literature (Steorts et al. 2014).

Task 1.2 comprises the binary classification of item pairs as duplicates or non-duplicates. In many contexts, this classification can be performed rather easily in a supervised manner. However, in the context of recommender systems, generally, no substantial amount of labeled training data (i.e., item pairs labelled as (non-)duplicates) is available for a supervised classification. Therefore, it is crucial to perform item pair classification in an unsupervised manner, not requiring any labeled training data (cf., e.g., Jurek et al. 2017). In the following, we describe the basic ideas of such an algorithm and emphasize the crucial peculiarities of the algorithm in the context of recommender systems. The algorithm starts with an initialization, followed by the proper classification and ends with all item pairs being classified as duplicate or non-duplicate.

The initialization consists of the selection of SMFs that are used for the classification. For each key attribute available in both data sets DS1 and DS2, adequate SMFs have to be specified. The choice of SMFs primarily depends on the data type of the respective key attribute. In particular, for key attributes containing string values and key attributes containing numerical values, different SMFs have to be used (e.g., the haversine SMF for GPS data values and the Jaro-Winkler SMF for string data values; cf. Table 1). Here, it is important to not only select one SMF per key attribute, but to select multiple SMFs with different characteristics, since the compared values of the key attributes may also exhibit varying deviations and specifications. For string attribute values with different suffixes (e.g., a restaurant is represented by "Fluffy's New York" in DS1 and by "Fluffy's Café & Pizzeria" in DS2), a SMF that focuses on the initial characters of a string such as the Jaro-Winkler SMF is appropriate. Further, for string attribute values with typographical errors (e.g., a restaurant is represented by "Fulffy's" in DS1 and by "Fluffys" in DS2), a SMF addressing this special deviation such as the Levenshtein SMF is suitable. Therefore, it is important to utilize multiple SMFs for item pair classification to cope with the challenges of highly diverse data values in the context of recommender systems. To further elaborate on the specification of SMFs for item pair classification, we give a broader discussion of selected SMFs with different characteristics in Table 1 based on Christen (2012) and state their properties and examples in the context of recommender systems.

The proper classification is then conducted via an unsupervised ensemble self-learning algorithm, which improves results compared to just using the values of SMFs for classification (Jurek et al. 2017). This self-learning algorithm starts with training a certain binary classifier. The training is conducted on a small set of training data, which consists of the item pairs with the highest similarity values (implicitly labeled as duplicates) and item pairs with the lowest similarity values (implicitly labeled as non-duplicates) and thus does not need to be labeled manually. This binary classifier is then used to predict the classes of all other item pairs. The item pairs classified with a high certainty are then added to the training data. Subsequently, the binary classifier is trained again and the steps are gradually repeated until all item pairs are classified as either duplicates or non-duplicates by this certain binary classifier. To further increase the robustness of the classification result, multiple such binary classifiers are used with the described self-learning method and the obtained results are then aggregated to obtain the final stable result of the item pair classification.

Table 1. Selected Similarity Measure Functions and their Application in the Context of Recommender Systems

Similarity measure functions	*Properties*	*Examples in the context of recommender systems*
Levenshtein The Levenshtein SMF is based on the minimum number of edit operations of single characters necessary to transform a string s_1 into a string s_2.	• Appropriate for misspellings/ typographical errors • Inappropriate for truncated/ shortened strings and divergent pre-/suffixes • Complexity: $O(\|s_1\| * \|s_2\|)$	Typographical error in the attribute "Restaurant Name": "Fulffy's" vs. "Fluffys".
Jaro The Jaro SMF is based on the number of agreeing characters c contained in the strings s_1 and s_2 within half the length of the longer string, and the number of transpositions t in the set of common substrings.	• Appropriate for misspellings/ typographical errors • Inappropriate for long divergent pre-/suffixes • Complexity: $O(\|s_1\| + \|s_2\|)$	Misspelling in the attribute "Restaurant Name": "Fluffy's Café" vs. "Flufy's Café".
Jaro-Winkler The Jaro-Winkler SMF extends the Jaro SMF, putting more emphasis on the beginning of the strings.	• Appropriate for misspellings/typographical errors and divergent suffixes • Inappropriate for long divergent prefixes • Complexity: $O(\|s_1\| + \|s_2\|)$	Divergent suffixes of the attribute "Restaurant Name": "Fluffy's New York" vs. "Fluffy's Café & Pizzeria".
Haversine This SMF is based on the haversine formula, which measures the distance between two locations on earth.	• Appropriate for geographical coordinates given in latitude/longitude	"40.711, -73.966" vs. "40.710, -73.965".

In **Task 1.3,** it is necessary to resolve multi-linkages of duplicates resulting from Task 1.2. This problem may arise as an item from DS1 can be contained in more than one item pair classified as a duplicate. Thus, this item from DS1 is linked to more than one item from DS2. Similarly, an item from DS2 can be linked to more than one item from DS1. As the matched items will be proposed to users in the recommendation step, it is important to resolve these multi-linkages of items to avoid redundant and multiple recommendations of individual items. To resolve the multi-linkages, the prediction scores of the ensemble classifier from Task 1.2 are used. Considering an item from DS1 linked to multiple items from DS2, only the linkage with the highest prediction score is retained and all other linkages are discarded. Analogously, only one linkage is kept when an item from DS2 is linked to multiple items from DS1. In this way, the n-to-n reference of items from DS1 and DS2 is firstly reduced to 1-to-n references and then to 1-to-1 references.

Step 1 concludes with the validation of the results of the duplicate detection in **Task 1.4**, which is necessary to assess the quality of the duplicate detection. This quality plays an important role in the context of recommender systems, as false duplicates would result in erroneous data integrations in the next step of the procedure, and thereby, to negative effects on item recommendations. On the other hand, false negatives would result in feasible data integrations not being carried out, thus reducing the benefit of the procedure. Therefore, a small excerpt of item pairs, serving as test

data, needs to be labeled as duplicates or non-duplicates for validation purposes. Here, a random selection of item pairs to be labeled would result in the vast majority of these item pairs being labeled as non-duplicates, since most item pairs are indeed non-duplicates. Therefore, it is important to take the calculated values of the SMFs into account and to also label item pairs which are more likely to be a real duplicate. Building on this labeled test data, the number of correct classifications (i.e., "true positives" and "true negatives") and the number of errors (i.e., "false positives" and "false negatives") can be determined. Based on these numbers, evaluation metrics such as precision, recall and F1-measure can be assessed. If these evaluation metrics report unsatisfactory results, the classification errors may be analyzed and tackled. The evaluation metrics thus enable to ensure a high quality of the conducted duplicate detection and to provide data suitable for the next step of the procedure, which concludes Task 1.4 and thus Step 1.

3.2 Data Integration in the Context of Recommender Systems

In Step 2 of the procedure, attributes and attribute values of DS1 and DS2 are integrated to obtain the envisioned more complete view on items. In particular, *matching* attributes (i.e., attributes of DS2 also existing in DS1) and *additional* attributes (i.e., attributes only existing in DS2) have to be identified and the items' attribute values have to be extended. To perform this integration in the context of recommender systems, we propose the following three Tasks 2.1-2.3 (cf. Fig. 4).

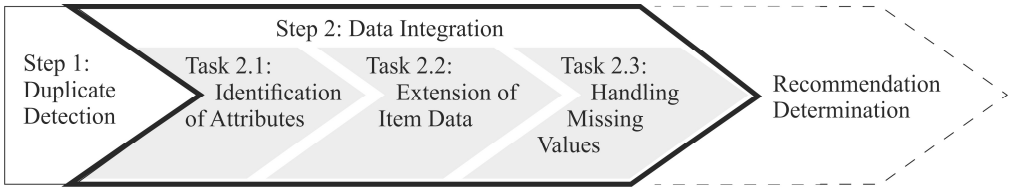

Fig. 4 The Step Data Integration in Detail

The goal of **Task 2.1** is to identify matching attributes. To do so, the attributes of DS2 have to be compared to the attributes of DS1. The automated identification of matching attributes can prove to be non-trivial in the context of recommender systems because different portals often use varying names for the same attribute (e.g., "Artist" and "Performer") due to heterogeneous data policies. An incorrect matching of attributes can lead to items being assigned wrong data and thus have a direct detrimental impact on recommendation quality. As this task is of relatively low complexity for humans, the identification may be performed in a manual manner (e.g., the manual matching of 143 attributes in DS1 to 251 attributes in DS2 in the application scenario regarding restaurants of our evaluation took approximately one hour and exhibited almost perfect inter-coder reliability). In contrast, an automated identification (e.g., using WordNet) may be error-prone, as it is difficult for an algorithm to directly identify attributes such as "Artist" and "Performer" as matching attributes. Furthermore, an automated identification requires a subsequent manual verification by humans, which is also time-consuming. Overall, an automated identification should only be performed when the number of attributes is extremely high, rendering a manual identification ineffective. In any case, all attributes of DS2 not matched to an attribute of DS1 are identified as additional attributes.

In **Task 2.2**, the item content data is extended for each item in DS1. More precisely, the item content data subsequently consists of the attributes of DS1 and the additional attributes of DS2. Additional attributes allow a more complete view on the considered item and may improve recommendation quality. In particular, additional attribute values can have enormous leverage for users with many item reviews in the context of recommender systems, since a large number of affected rated items can be described in more detail with the additional content. Depending on the recommender system used or under trade-off considerations, it may be helpful to limit the number of the additional attributes considered for data extension. To identify a subset of additional attributes for which a strong improvement of recommendation quality is expected (e.g., attributes with very many missing values may hardly impact recommendation quality), several options are possible (e.g., the use of an attribute selection algorithm; cf. Chandrashekar and Sahin 2014; Molina et al. 2002). These options are discussed in more detail in Section 4.3. After selecting the additional attributes, for each item in DS1 for which a duplicate in DS2 was identified and for each additional attribute chosen, the respective attribute values of the duplicate are inserted into the item content data.

After Task 2.2, some attribute values of items in the extended data set may still be missing because they are not provided by either data set (e.g., the values of the attribute "Genres" are not given for all items in the movie domain). These missing values have to be addressed in **Task 2.3**, since many recommender systems cannot operate on missing attribute values. Moreover, missing attribute values may be detrimental to recommendation quality. Therefore, a further extension of item content data is enabled by imputation methods. More precisely, missing attribute values can be inferred via imputation based on non-missing attribute values in the extended data set. Here, the presented

procedure provides an advantage compared to imputing values based on just DS1 as the attribute values from both data sets DS1 and DS2 are available and can be used as basis for the imputation. Table 2 discusses selected imputation methods and their relevance in the context of recommender systems based on Enders (2010). In addition to these imputation methods, it is also feasible to impute values in a user-specific way which is more flexible than assigning fixed values for the missing values in the extended data set. In this case, the missing values of all items rated by a user can be handled by an imputation method from Table 2 (e.g., Arithmetic Mean Imputation) to capture the user's preferences more accurately when generating her/his user profile.

Table 2. Selected Methods for Handling Missing Values and their Application in the Context of Recommender Systems

Imputation methods	*Properties*	*Examples in the context of recommender systems*
Arithmetic Mean Imputation (AMI) Missing attribute values are replaced with the mean attribute value of all items, where the values for this attribute are not missing.	• AMI is convenient to implement • AMI attenuates standard deviation and variance	Each missing value of the attribute "Runtime" is replaced with the mean value of "Runtime" (as an indicator) over all movies that do have a value for "Runtime".
Regression Imputation (RI) Missing values are replaced with predicted scores from regression equations. The regression equations are estimated by analyzing the extended data set.	• RI is complicated to implement • RI attenuates standard deviation and variance (but less than AMI)	For two hotel attributes "Price" (P_i) and "Service" (S_i), there are only missing values for "Service". A regression equation $\hat{S}_i = \hat{\beta}_0 + \hat{\beta}_1(P_i)$ for the attribute "Service", depending on the attribute "Price", is estimated by analyzing the hotels with given values for "Service". The missing values S_i of "Service" are replaced by $\hat{S}_i$.
Hot Deck Imputation (HDI) Missing attribute values of an item are replaced with the corresponding values of the most similar item.	• HDI is convenient to implement • HDI attenuates standard deviation and variance (but less than AMI)	The movie "The Dark Knight" is the most similar movie to "The Dark Knight Rises", as both movies belong to the batman trilogy of the director "Christopher Nolan". The value of "The Dark Knight" for the attribute "Genres" is "Action" and thus, the missing value of "The Dark Knight Rises" for "Genres" is inferred with the value "Action".

3.3 Subsequent Step: Recommendation Determination

Subsequent to duplicate detection and data integration, recommendations for users on e-commerce platforms can be inferred by applying a recommender system based on the extended data set and evaluating the resulting recommendations. This step is also necessary to analyze the effects of data set extension on recommendation quality. As our approach is tailored to data sets containing item content data in addition to rating data, it is feasible to apply both content-based as well as hybrid recommender systems that leverage both data types (Ricci et al. 2015b). Handling item content data is very important in e-commerce settings, because the recommender system can map the potentially extensive needs of customers more accurately due to the more precise description of the items (e.g., proposal of tailored products based on product preferences). Therefore, for this *subsequent* step of our procedure, we suggest to apply the state-of-the-art hybrid recommender system approach Content-Boosted Matrix Factorization (CBMF; cf. Forbes and Zhu 2011), which utilizes both rating data and, in particular, item content data and is thus more comprehensive than collaborative filtering recommender systems. Matrix factorization approaches have become very popular through the Netflix contest, which started in 2006 and ended in 2009 (Koren 2009; Koren et al. 2009), and now constitute state-of-the-art recommender systems (Kim et al. 2016; Ning et al. 2017). As a matrix factorization approach, CBMF learns a model by optimizing a loss function based on training data and therefore, preliminary steps such as attribute weighting or attribute selection are not necessary for CBMF (Koren 2009; Nguyen and Zhu 2013).
The basic idea of matrix factorization recommender systems is to decompose the rating matrix R (users as rows; items as columns) into two low-rank matrices P (representing users) and Q (representing items), with $PQ \approx R$. Then, the

idea of CBMF is to further decompose the matrix Q into a low-rank matrix A and the matrix F, with $AF^T = Q$ and F containing the attribute vectors of items (items as rows; attributes as columns). Hence, the overall idea is that the rating matrix R can be approximated by $R \approx PAF^T$. In particular, CBMF learns a n-dimensional vector of latent factors $p_u \in \mathbb{R}^n$ for each user u and a n-dimensional vector of latent factors $a_f \in \mathbb{R}^n$ for each attribute f, such that the actual rating r_{ui} for a user-item pair (u, i) is approximated by the predicted star rating $\hat{r}_{ui} = p_u^T q_i$, with $q_i = \sum_{f \in F_i} a_f$ and F_i being the set of attributes that are assigned to the item i. Finally, to evaluate the effects of the data set extension on recommendation quality, the rating data is split into training data for learning the parameters of the CBMF model (p_u and a_f) and test data to assess the recommendation quality via quality measures such as Root-Mean-Square-Error (RMSE; cf. Shani and Gunawardana 2011).

4 Evaluating the Procedure in Real-world Scenarios

In this section, we evaluate the proposed procedure in two real-world e-commerce scenarios. First, the reasons for selecting these scenarios are discussed and the used data sets are described. Thereafter, the evaluation of the procedure with respect to these data sets is outlined. Finally, important effects of the data set extension regarding items, content and users on recommendation quality are presented.

4.1 Selection and Description of the Real-world Scenarios

We evaluated the procedure in two real-world e-commerce scenarios regarding the domains of restaurants and movies. While these domains are frequent subjects of IS research in e-commerce (Chang and Jung 2017; Nguyen et al. 2018; Wei et al. 2013; Yan et al. 2015), both domains exhibit versatile facets and different challenges for a procedure for data set extension. Thereby, analyzing these two domains allows for a broader evaluation of the proposed procedure in e-commerce application scenarios.

First, we selected the domain of restaurants because this domain is very challenging regarding duplicate detection (i.e., Step 1 of the procedure, e.g., the resolution of multi-linkages of duplicates (Task 1.3)) in the context of recommender systems. In comparison to other domains (e.g., the domain of movies as second scenario) there are items with the same name being found in the immediate vicinity (i.e., in the case of restaurant chains such as McDonald's or Subway), which makes this domain especially challenging. For the real-world scenario in the domain of restaurants, we prepared data sets of two leading advertising web portals which provide crowd-sourced ratings about businesses (e.g., restaurants). The first portal (Portal R1) focuses on travel opportunities and businesses such as restaurants and provided over 650 million ratings whereas the second portal (Portal R2) specializes on local businesses such as bars or restaurants and provided over 150 million ratings by 2020. These portals were chosen because an initial check revealed that, while both portals contain data about an overlapping set of real-world entities, they offer an interestingly different view (i.e., different attributes) on these entities. In particular, we selected the area of New York City (USA) as both portals provided a large number of items, users and ratings for this area. In this way, the evaluation of the procedure and the analysis regarding its effects on recommendation quality could be performed on a sufficiently large data basis. Here, the data from Portal R1 consists of more than 8,900 items representing restaurants in the area of New York City, rated by over 380,000 users with approximately 850,000 ratings. The data from Portal R2 consists of over

Table 3. Description of the Restaurant Data Sets

	Portal R1 (DS_{R1})	**Portal R2 (DS_{R2})**
# of items (restaurants)	8,909	18,507
# of users	386,958	583,815
# of ratings	855,357	2,396,643
# of key attributes	6	6
# of further attributes (category attributes and business information attributes)	143	251
# of possible attribute values	1,247,260	4,589,736
# of missing values	3,253 (0.26%)	190,789 (4.16%)

18,500 items representing restaurants in the same area, rated by more than 580,000 users with around 2.4 million ratings. Each item of Portal R1 is described by the key attributes "Name", "Postal Code", "Geolocation", "Address", "Phone" and "District", category attributes such as "Italian Cuisine" or "Pizza", and business information attributes

such as “Parking Available” or “Waiter Service”. In Portal R2, items are described by key attributes equaling those in Portal R1 as well as (partly different) category attributes and business information attributes. The data from Portal R1 contains around 3,000 missing values for one attribute whereas the data from Portal R2 contains more than 190,000 missing values for 26 attributes. In our evaluation, we extended the data from Portal R1 with the data from Portal R2 (i.e., the data from Portal R1 served as DS_{R1} and the data from Portal R2 served as DS_{R2}). Table 3 describes the restaurant data sets.

In addition, we selected the domain of movies because this domain exhibits further but different challenges regarding item content data extension in the context of recommender systems. In comparison to the restaurant domain, the detection of duplicates and in particular the resolution of multi-linkages of duplicates is less challenging in the movie domain, since different movies have usually different titles (as key attribute) due to copyright standards. Nevertheless, Step 1 of the procedure is still favorable for movies in order to detect non-trivial movie duplicates in case the movie titles do not exactly match, as key attributes can (slightly) vary between different portals in some cases (e.g., the movie titles “Mission: Impossible – Ghost Protocol” and “Mission: Impossible – Ghost Protocol (2011)” represent the same item). Moreover, an initial check revealed that the amount of missing values in the data sets of both movie web portals (Portal M1 and Portal M2) is very high compared to other domains (e.g., restaurants). This means that Step 2 of the procedure including the task of handling missing values is even more important for the real-world scenario in the movie domain. Hence, we prepared data sets of two leading web portals which provide crowd-sourced ratings about movies. Here, the data from Portal M1 consists of approximately 29,000 movie items, rated by over 425,000 users with nearly 530,000 ratings. The data from Portal M2 consists of over 12,500 movie items, rated by approximately 230,000 users with nearly 410,000 ratings. Each item of Portal M1 is described by the key attribute “Title” and further attributes such as “Brand”. In Portal M2, items are described by the same key attribute as in Portal M1 as well as by further attributes such as “Cast” and “Language”. The data from Portal M1 contains over 245,000 missing values for all attributes whereas the data from Portal M2 contains more than 1 million missing values for all attributes. In our evaluation, we extended the data from Portal M1 with the data from Portal M2 (i.e., the data from Portal M1 served as DS_{M1} and the data from Portal M2 served as DS_{M2}). Table 4 describes the movie data sets.

Table 4. Description of the Movie Data Sets

	Portal M1 (DS_{M1})	**Portal M2 (DS_{M2})**
# of items (movies)	28,973	12,842
# of users	428,519	230,151
# of ratings	528,777	409,935
# of key attributes	1	1
# of further attributes	13	103
# of possible attribute values	376,649	1,322,726
# of missing values	247,341 (65.67%)	1,082,387 (81.83%)

4.2 Evaluation of the Procedure

In this section, we discuss the evaluation of the procedure for extending data sets with item content data in the restaurant and movie domain and present the evaluation results for each step for both domains.

Evaluation of Step 1 – Duplicate Detection

In the following, we outline the evaluation of the duplicate detection step. More precisely, the goal of this section is to assess the evaluation metrics precision, recall and F1-measure of duplicate detection. Therefore, we first discuss how we conducted and validated the tasks of this step and then present the evaluation results.

Since this step is more challenging for restaurants, we especially focus on this domain.

To begin with, in Task 1.1, the key attribute values (cf. Table 5) of DS_{R1} and DS_{R2} were standardized due to inconsistent values caused by heterogeneous data policies among restaurant portals. For example, the postal code in DS_{R1} was given in the format “ZIP+4” (containing the standard five-digit postal code with four additional digits for postal delivery, e.g., “10019-2132”) and in DS_{R2} in the format “ZIP” (containing the standard five-digit postal code, e.g., “10019”). Hence, “Postal Code” was restricted to only the standard five-digit postal code “ZIP” (e.g., “10019”) to achieve standardized key attribute values. In the data preparation subtask, pairs of restaurants which were more than 1,000 meters apart from each other based on the key attribute “Geolocation” were removed, due to these restaurant pairs being obvious non-duplicates. This led to a total of 11,492 item pairs, constituting the data for the next

task “Item Pair Classification”. Task 1.2 was initialized by selecting adequate SMFs for all key attributes, following the argumentations given in Section 3. For example, the SMFs “Jaro-Winkler” and “Levenshtein” were proved as useful for the key attributes “Name” and “Address” and the SMF “Haversine” was beneficial for “Geolocation” (Kamath et al. 2013). These key attributes were selected as no natural unique IDs for the restaurants were available across DS_{R1} and DS_{R2}. The duplicate detection then yielded at first 6,226 pairs classified as duplicates and 5,266 item pairs classified as non-duplicates. In Task 1.3, multi-linkages of items were resolved. For example, the restaurant “Sushi You” in DS_{R1} was contained in two item pairs classified as duplicates (with the restaurant “Sushi You” from DS_{R2} in the first pair and with the restaurant “Sushi Ko” from DS_{R2} in the second pair). Here, the prediction score of the first pair was higher than the score of the second one and therefore, only the first pair was retained. After resolving such multi-linkages, the number of duplicate item pairs decreased to 5,919. With regard to Task 1.4, 500 item pairs (250 items presumed to be duplicates and 250 items presumed to be non-duplicates) were selected to validate our duplicate detection step. Thereby, the item pairs were examined by a web-based search which involved 1) visiting the homepages of the restaurants, 2) searching the restaurants via *Google Maps* and 3) using *Google Street View* to check the identity of restaurants. This method was necessary to reliably determine actual duplicates and non-duplicates as some non-duplicate item pairs were hard to identify. For example, the restaurants “Murray’s Cheese Shop” in DS_{R1} located at “254 Bleecker St” in “West Village” and “Murray’s Cheese Bar” in DS_{R2} at “264 Bleecker St” in “West Village”, which seem to be very similar at first sight, turned out to be non-duplicates after the examination. The validation of the duplicate detection yielded a precision of 95.9% (i.e., 235 of 245 classified duplicates were real duplicates; 240 of 255 classified non-duplicates were real non-duplicates), a recall of 94.0% (i.e., 235 of 250 real duplicates were classified as duplicates; 240 of 250 real non-duplicates were classified as non-duplicates) and a F1-measure of 94.9%, demonstrating a very high quality. Summing up, the first step of the procedure yielded 5,919 duplicate restaurant item pairs of high quality constituting the basis for Step 2 of the procedure.

Table 5. Key Attributes of both Restaurant Portals

Key attributes	*Data type*	*Example key attribute values from both portals for a duplicate*
Name	String	“9 Ten Restaurant” (in DS_{R1}), “9 10 Restaurant” (in DS_{R2})
Postal Code	Number	“10019-2132” (in DS_{R1}), “10019” (in DS_{R2})
Geolocation	Geographic coordinates (latitude and longitude)	“N 40.76591° / W -73.97979°” (in DS_{R1}), “N 40.7659964050293° / W -73.9797178100586°” (in DS_{R2})
Address	String	“910 Seventh Avenue” (in DS_{R1}), “910 7th Av” (in DS_{R2})
Phone	Number	“+1 917-639-3366” (in DS_{R1}), “(917) 639 3666” (in DS_{R2})
District	String	“Midtown” (in DS_{R1}), “Midtown West” (in DS_{R2})

Next, we briefly outline the first step of the procedure for the movie domain. As described before, the duplicate detection step for the movie domain is in general less challenging than for the restaurant domain due to copyright standards. However, titles of movie duplicates do not always exactly match, since different movie portals have heterogeneous data policies (e.g., the movie titles “Mission: Impossible – Ghost Protocol” and “Mission: Impossible – Ghost Protocol (2011)” represent the same item). Hence, standardization of the key attribute “Title” in both data sets DS_{M1} and DS_{M2} is necessary (e.g., removing the year of the movie’s release). Thereafter, many duplicates can be detected directly by matching the standardized “Title” of movies in a large part of the cases (cf. Section 4.1). Similar as for restaurants, pairs of movies which were obvious non-duplicates (based on similarities of the key attribute “Title”) were removed during blocking leading to 10,160 item pairs as result of Task 1.1. Since DS_{M1} also contained items going beyond regular cinematographic movies (e.g., other film material such as “The Theory of Evolution: A History of Controversy”), item pairs could only be identified for the mentioned 10,160 items in DS_{M1}. In Task 1.2, SMFs such as “Jaro-Winkler” and “Levenshtein” were used for the key attribute “Title” for conducting item pair classification similarly as for restaurants. With no multi-linkages present in the result of Task 1.2 (i.e., Task 1.3 could be skipped), 9,438 movie item pairs were detected as duplicates. Similarly, as for restaurants, 500 item pairs were prepared to validate duplicate detection by a manual web-based search. The validation of the duplicate detection for movies in Task 1.4 yielded a precision of 95.1%, a recall of 96.7% and a F1-measure of 95.9%, demonstrating a very

high quality for detecting duplicates. Summing up, the first step of the procedure yielded 9,438 duplicate movie item pairs of high quality constituting the basis for Step 2 of the procedure.

Evaluation of Step 2 – Data Integration

In this section, we outline the evaluation of the data integration step. The goal of this section is to assess how the completeness of the item content data could be increased through data integration. Therefore, we first establish how we conducted and validated the tasks of Step 2 of the procedure and then present the results of the evaluation. Since the number of further attributes in DS_{M2} (compared to DS_{M1}) and the numbers of missing attribute values in DS_{M1} and DS_{M2} are very high (cf. Table 4), Step 2 is of particular relevance for the real-world scenario regarding the movie domain. Nevertheless, Step 2 is also crucial for the real-world scenario regarding restaurants, as in this step the actual data set extension is performed.

Following Task 2.1, as heterogeneous data policies among portals in the restaurant domain had led to different names of the same attribute and different levels of granularity used across DS_{R1} and DS_{R2}, all attributes of DS_{R2} were compared to the attributes of DS_{R1} to identify matching and additional attributes. Thereby, 57 attributes of DS_{R2} such as "Japanese", "Pizza" or "Vegan" were identified as matching attributes and 194 attributes of DS_{R2} such as "Attire", "Karaoke" or "Take Out" were identified as additional attributes in a manual check requiring approximately one hour of work, exhibiting almost perfect inter-coder reliability. According to Task 2.2, these additional attributes are to be analyzed regarding an extension of DS_{R1}. Here, for a first evaluation regarding the effects on recommendation quality, we used all additional attributes for the extension of DS_{R1}. Thus, the extended data set contained all attributes of DS_{R1} and all additional attributes of DS_{R2}. Thereafter, the item content data of DS_{R1} was extended and attribute values of duplicates were inserted. Further, we validated Task 2.3, which means, the remaining missing attribute values were imputed in a first step. To this end, we evaluated the use of the Hot Deck Imputation method (cf. Table 2), allowing the replacement of all missing values and yielding an item content data set without missing values. In total, the evaluation shows that the completeness of the item content data of DS_{R1} can be increased by integrating 194 additional attributes from DS_{R2} and by imputation of 3,253 values in DS_{R1} and 190,789 values in DS_{R2}. This emphasizes the superior data quality of the resulting extended data set compared to the basis data set DS_{R1} regarding the dimension completeness.

In the case of the movie data sets, all 103 attributes of DS_{M2} such as "Genres", "Cast" or "Language" were identified as additional attributes in Task 2.1. In Task 2.2, for a first evaluation regarding the effects on recommendation quality, we used all additional attributes of DS_{M2} for the extension of DS_{M1} similar to the case of restaurants. Thus, the attributes and values were inserted for the identified duplicates and thus, the extended data set contained all attributes of DS_{M1} and all attributes of DS_{M2}. In Task 2.3, the remaining missing attribute values were imputed by means of the Hot Deck Imputation method (cf. Table 2) yielding an item content data set without missing values. In total, the evaluation shows that the completeness of the item content data of DS_{M1} can be increased by integrating 103 additional attributes from DS_{M2} and by imputation of 247,341 values in DS_{M1} and 1,082,387 values in DS_{M2}. Therefore, the resulting extended data set shows strongly increased data quality compared to the basis data set DS_{M1} regarding the dimension completeness.

Evaluation of Subsequent Step – Recommendation Determination

Finally, we discuss the evaluation of the recommendation determination based on the extended data sets with increased completeness regarding both domains. After the data set extension in the first two steps of the procedure, the recommendations based on the extended data sets could be computed. As indicated in Section 3, we validated whether the hybrid recommender system approach CBMF (Forbes and Zhu 2011; Nguyen and Zhu 2013) can be utilized. We followed Nguyen and Zhu (2013) in regard to the default configuration for CBMF, with the only exception being the regularization penalty factor λ, which has to be adjusted depending on the data set at hand (Koren et al. 2009). To this end, we compared the results of cross-validation tests of different values for λ as described by Koren et al. (2009). In these tests, the value $\lambda = 10^{-5}$ yielded the best results in terms of RMSE. After the execution of CBMF, the recommendations were evaluated by the following standard technique (cf., e.g., Shani and Gunawardana 2011). The ratings of DS_{R1} and DS_{M1} were randomly split into a training set (67% of ratings) to learn the parameters of the CBMF model (p_u and a_f, cf. Section 3) and a test set (33% of ratings) for assessing recommendation quality. We quantified recommendation quality by the RMSE between the real ratings and the predicted ratings of the CBMF in the test set. To assess the recommendation quality based on the extended data sets compared to just data sets DS_{R1} or DS_{M1}, respectively, the training of the CBMF parameters and the assessment of recommendation quality were validated on either the item content data of the extended data set or just on the item content data of DS_{R1} or DS_{M1}. Here, in both cases (extended data set compared to the basis data set) the train-test-split remained the same such that a meaningful comparison of both cases was possible for both domains. The recommendation determination could be applied in each

case without restrictions and yielded recommendations for each user. In particular, our procedure was able to successfully navigate numerous challenges in this context (cf. Table 6), which are common when trying to extend an item content data set with respect to the data quality dimension completeness. This successful validation of the determined recommendations concludes the evaluation of the proposed procedure in both real-world scenarios.

Table 6. Challenges in the Context of Recommender Systems

Topics	*Challenges in the context of recommender systems*	*References to procedure step / task*
Data / Content	• Decentral data capturing by many different users results in data quality problems requiring standardization • Heterogeneous data policies among portals lead to different characteristics of the data across data sets, also requiring standardization • Item content data is a central decisive factor for e-commerce business models and respective recommender systems	**1.1 Data Standardization and Preparation**
Key Attributes and Item Pair Classification	• Labeled training data is missing in the context of recommender systems for a supervised item pair classification • No natural unique IDs are available for items (e.g. restaurants) • Values of key attributes are entered in a decentral way and depend on the users' own interpretation leading to highly diverse data values • Items with the same name referring to the same organization (e.g., "McDonald's") and items with similar names referring to different organizations (e.g., "Sushi You" vs. "Sushi Ko") in the restaurant domain are potentially in close proximity in urban areas; however, they have to be distinguished as separate items	**1.2 Item Pair Classification**
Matching Attributes	• Heterogeneous data policies among portals lead to different names of the same attribute (e.g., "Bar" vs. "Pub") • Portals potentially use different levels of granularity when describing the attributes (e.g., "Asian Cuisine" vs. "Japanese Cuisine")	**2.1 Identification of Attributes**
Additional Attributes	• Attributes and their values (e.g., eight times more attributes after data set extension in the movie domain) directly affect the quality of the recommender system and the resulting recommendations	**2.2 Extension of Item Data**
Missing Values	• Many recommender system techniques cannot handle missing values (e.g., 75% missing attribute values had to be imputed in the movie domain)	**2.3 Handling Missing Values**

4.3 Effects on Recommendation Quality

In addition to the evaluation of the procedure itself in Section 4.2, we observed and examined effects of our procedure on recommendation quality in both e-commerce real-world scenarios. These effects can serve as a starting point for further investigations of the impact of completeness on the recommendation quality based on our procedure (cf. Section 2.2). In particular, besides evaluating the general impact of increased completeness on recommendation quality when applying the proposed procedure (Effect 1), we also investigated effects in detail on the results of the procedure from the three major dimensions related to (content-based and hybrid) recommendations in e-commerce (Heinrich et al. 2019): Items (Effect 2), content in form of attributes (Effect 3) and attribute values (Effect 4), and users (Effect 5). An overview of the results regarding these effects for both the restaurant and the movie domain is given in Table 7.

Effect 1. Extending the basis data set (DS_{R1} and DS_{M1}, respectively) by applying the proposed procedure improved recommendation quality considerably.

Scenario regarding restaurants: Indeed, the more complete view on restaurants provided by the extended data set led to an improvement in recommendation quality of 13.2% (the RMSE achieved for the extended data set is 0.89, while the RMSE for just DS_{R1} is 1.02). The more complete view and its effect can be illustrated by an example considering the user "Michelle", who had submitted 43 ratings overall. This user had, in reality, rated the restaurant "ShunLee" with a score of 4 stars. The rating of this restaurant as estimated by CBMF based on just DS_{R1} was 1 star, which means that there was a huge discrepancy between the real and the estimated rating. In the extended data set, the item vector of "ShunLee" was extended by all additional attributes and attribute values of its duplicate in DS_{R2} as described above. This extension led to a large improvement, as CBMF based on the extended data set determined a rating of 3 stars, which is much closer to the real rating of the user. Overall, the recommendations for "Michelle" based on the extended data set and based on just DS_{R1} resulted in RMSEs of 0.56 and 3.78, respectively. This example further illustrates the (considerable) improvement of recommendation quality.

Scenario regarding movies: Compared to the restaurant domain, the overall effect of the procedure in the movie domain is even stronger, as the extension of DS_{M1} led to an improvement in recommendation quality of 24.6%. However, the baseline RMSE of 3.15 based on just DS_{M1} is inferior for the movie domain compared to the restaurant domain with a baseline RMSE of 1.02, which means, improving a higher baseline RMSE is comparatively easier. This puts the high improvement in recommendation quality in perspective. Besides this, individual analyses of users regarding improvements in recommendation quality can be performed analogously to the description above for restaurants.

Effect 2. A sophisticated duplicate detection as proposed by our procedure yielded a high improvement in recommendation quality.

Scenario regarding restaurants: In order to investigate the importance of duplicate detection (cf. Section 3.1) on the resulting recommendation quality, we further instantiated and evaluated the procedure with an alternative rule-based duplicate detection algorithm (cf. Christen 2012). To evaluate this alternative algorithm, we performed Task 1.1, Task 1.3 and Task 1.4 in the same way, but for Task 1.2, we chose the following decision-rule aiming for a simple but transparent classification of item pairs (A, B):

If $jaro\_winkler\_similarity_{name}(A,B) > T_1$ **and** $haversine\_similarity_{geolocation}(A,B) > T_2$ **then** item B is classified as a duplicate of item A **else** item B is not classified as a duplicate of item A.

We evaluated different threshold configurations for T_1 and T_2 resulting in the best validation results for the thresholds $T_1 = 0.9$ and $T_2 = 0.909$ (corresponding to a distance of 100 meters), which were used for the rule-based item pair classification. As the rule-based duplicate detection was rather restrictive with judging pairs of items to be a duplicate, the fewer pairs of items identified as duplicates by the rule-based duplicate detection were almost all correctly classified, resulting in a high precision of 96.8% (compared to 95.9% precision of the sophisticated duplicate detection). However, the rule-based duplicate detection mainly just identified the rather obvious duplicates, leading to this high precision but a quite low recall. More precisely, it was only able to identify 72.8% of duplicates as indicated by the recall (compared to 94.0% recall of the sophisticated duplicate detection). Thus, the rule-based duplicate detection also exhibited an overall lower F1-measure of 83.1% compared to 94.9% for the sophisticated duplicate detection, demonstrating the higher quality of the sophisticated duplicate detection. The assessed improvement in recommendation quality when conducting the remainder of the procedure using this duplicate detection with lower quality was only 9.8% (compared to 13.2% improvement for the sophisticated duplicate detection with higher quality assessed on the same test set of ratings as in Effect 1). These results show that the sophisticated duplicate detection algorithm proposed by our procedure led to a significantly higher improvement in recommendation quality.

Scenario regarding movies: Similarly, as for restaurants, we instantiated and evaluated a rule-based duplicate detection algorithm in the movie domain yielding 85.3% for F1-measure (compared to 95.9% for the sophisticated duplicate detection). Nevertheless, even the procedure with the rule-based duplicate detection yields an improvement in recommendation quality by 23.9%, which is smaller than the improvement based on the sophisticated duplicate detection, which is 24.6%.

Effect 3. The extension of the basis data set (DS_{R1} and DS_{M1}, respectively) with further attributes (of DS_{R2} and DS_{M2}, respectively) generally supported the increase in recommendation quality, with the extent of improvement depending on the attribute set used for the extension.

Scenario regarding restaurants: To analyze and separate the effect of additional attributes for extension in Task 2.2, we split all additional attributes from DS_{R2} into two equally sized groups based on the absolute number of available values per attribute. First, we extended DS_{R1} with the set of additional attributes from DS_{R2} with a low number of available attribute values (Set 1), leading to an improvement in recommendation quality of just 0.1%. Second, the extension of DS_{R1} with the set of additional attributes with a high number of available attribute values (Set 2) achieved

an improvement of 12.6%. In comparison, the extension of DS_{R1} with all additional attributes of DS_{R2} (Set 3) led to an improvement of 12.7%.[4] These results show that while the extension with additional attributes generally contributed to an improvement of recommendation quality, the extent of improvement depended on the number of available attribute values of the additional attributes. Thus, these results indicate that the increase in recommendation quality could mainly be traced back to attributes with a high number of available attribute values. Moreover, we investigated the extension of DS_{R1} with *all attributes* of DS_{R2} (Set 4; i.e., additional attributes *and* matching attributes from DS_{R2}) in order to further analyze this effect. This means, we omitted the identification of matching attributes (cf. Task 2.1) and extended DS_{R1} with all attributes of DS_{R2} (i.e., additional and matching attributes). Although another 57 (matching) attributes were added compared to the extension with only additional attributes, the improvement of recommendation quality decreased slightly by 0.1% to 12.6%. This finding based on our chosen real-world scenario supports that more data (i.e., more attributes and attribute values) does not always lead to better results of decision support systems and, in particular, recommender systems (cf. Section 2.2). Therefore, the additional and more complete data provided by the matching attributes did not yield any added value, which is in line with works such as Bleiholder and Naumann (2008). In our application context, the matching of attributes led to just a slight improvement of the recommendation quality (0.1%), however, there may be application areas in which the matching of attributes contributes even more to an improvement of the recommendation quality and therefore Task 2.1 of the procedure is essential.

Since both adding attributes and identifying matching attributes may cause effort, it would be interesting to further investigate how to choose an adequate balance between these efforts and the resulting benefits of improved recommendation quality. For instance, when the efforts for adding attributes are low, all additional attributes can be selected for extension. Otherwise, a limitation to a smaller set of (additional) attributes (e.g., attributes with a high number of available attribute values) may be reasonable to reduce high efforts while simultaneously accomplishing a similarly high improvement of recommendation quality.

Scenario regarding movies: As for restaurants, we analyzed four sets of additional attributes (Set 1-4) from DS_{M2} regarding an improvement in recommendation quality. Since the scenario regarding movies did not yield matching attributes, all attributes of DS_{M2} constituted additional attributes and thus, the attribute sets Set 3 and Set 4 were identical. Here, the results regarding this effect for movies further underline the findings identified for restaurants as the improvement of 1.7% in recommendation quality for Set 1 was small compared to high improvements of 17.4% for the Sets 2-4. That is, the increase in recommendation quality could mainly be traced back to attributes with a high number of available attribute values.

Effect 4. More attribute values (i.e., less missing values) resulted in increased recommendation quality.

Scenario regarding restaurants: In addition to the analysis of the set of attributes, we also investigated effects of item content data with respect to (missing) attribute values. We fixed the set of attributes in the extended data set and focused on the imputation of missing attribute values (cf. Task 2.3) in order to separate Effect 4. We examined three settings with a varying number of (missing) attribute values. In the first setting, we imputed all missing values according to Task 2.3, resulting in no missing values in the item content data set used. The second setting used the extended data set without imputing missing values. In our real-world scenario regarding restaurants, however, only four percent of attribute values were missing, which could limit the extent of potential effects of missing attribute values. Therefore, we considered a third setting, in which we randomly removed an additional ten percent of attribute values from the extended item content data set to examine the effect of missing attribute values more generally in the restaurant domain. This led to a total of fourteen percent of missing attribute values in this third setting. We evaluated all three settings regarding resulting improvements in recommendation quality (i.e., RMSE based on the extended data set vs. RMSE based on just DS_{R1}). The results showed an improvement in recommendation quality of 13.2% for the first setting, 12.7% for the second setting and 6.5% for the third setting.

Scenario regarding movies: In contrast to the scenario regarding restaurants, the movie data sets showed high numbers of missing attribute values (cf. Table 4) making this scenario especially promising for analyzing the effect of imputing missing values (in Step 2 of the procedure) on recommendation quality in a real-world e-commerce application scenario. Similarly, as for restaurants, we examined the three settings with a varying number of missing attribute values. The results showed an improvement in recommendation quality of 24.6% for the first setting (i.e., the extended data set with imputed missing values), 17.4% for the second setting (i.e., the extended data set without imputed missing values) and 13.7% for the third setting (i.e., the extended data set without imputed missing values and 10% further removed attribute values).

[4] The difference between the improvement of 12.7% in Effect 3 and the improvement of 13.2% in Effect 1 can be attributed to the fact that imputation of missing values is omitted in Effect 3.

These results emphasize that recommendation quality benefits significantly from having more attribute values and, in particular, from imputing missing values, which constitutes a main task in the proposed procedure (cf. Task 2.3).

Effect 5. Users with a high number of submitted ratings benefitted more from the data set extension than users with a low number of submitted ratings.

Scenario regarding restaurants: For the analysis of this effect, we examined the relation between the number of ratings submitted by users and the increase in recommendation quality. To do so, we grouped all users into three equally sized groups based on their number of submitted ratings in the training set and examined the three groups individually regarding their improvement in recommendation quality. The first group containing users with the highest number of ratings (averaging about 29 ratings submitted per user) achieved a RMSE improvement of 17.1%. The second group, whose users had on average submitted about 15 ratings, recorded a RMSE improvement of 16.3%. Finally, the third group of users, with an average of about 10 ratings submitted per user, achieved the lowest improvement of recommendation quality, accomplishing a RMSE improvement of 9.9%.

Scenario regarding movies: Analogous as for restaurants, we grouped the users in the movie scenario into three equally sized groups. The first group, whose users had on average submitted about 4 ratings, achieved the highest RMSE improvement of 45.4%. The second group, whose users had submitted about 2 ratings on average, still recorded a high RMSE improvement of 42.7%. Finally, the third group of users, with an average of about 1 rating submitted per user, achieved the lowest improvement of recommendation quality, accomplishing a RMSE improvement of only 6.0%. Although the improvement for the third user group is small, it is still noteworthy as these users with just 1 submitted rating have only rating data in either the training set or the test set. In particular, this means that even users without ratings at all (i.e., without ratings in the training set) benefit from extending the item content data set, which is of high relevance for e-commerce applications, as the case of new users occurs very frequently.

Overall, these results indicate that the improvement of recommendation quality depended on the number of ratings submitted by users, and that users with a higher number of submitted ratings benefitted more. In a detailed analysis, we concluded that this effect can be attributed to the fact that users with a higher number of submitted ratings mainly rated items for whom more item content was added. Thus, the extended data set enabled the recommender system to infer these users' ratings even more accurately.

Table 7. Overview of Improvements in Recommendation Quality for each Effect

Effects	*Evaluation configurations*		*Relative improvements in recommendation quality (RMSE) by procedure application*	
			Restaurants	*Movies*
1	Standard procedure configuration (as outlined in section 4.2)		13.2%	24.6%
2	Procedure with simplified rule-based duplicate detection		9.8%	23.9%
3	Procedure without imputation and …	additional attributes with low number of available attribute values (Set 1)	0.1%	1.7%
		additional attributes with high number of available attribute values (Set 2)	12.6%	17.4%
		all additional attributes (Set 3)	12.7%	17.4%
		all attributes of DS2 (Set 4)	12.6%	17.4%
4	Standard procedure configuration (as outlined in section 4.2) (Setting 1)		13.2%	24.6%
	Procedure without imputation (Setting 2)		12.7%	17.4%
	Procedure without imputation and further removed attribute values (Setting 3)		6.5%	13.7%
5	Procedure for users with high rating numbers (Group 1)		17.1%	45.4%
	Procedure for users with moderate rating numbers (Group 2)		16.3%	42.7%
	Procedure for users with low rating numbers (Group 3)		9.9%	6.0%

5 Conclusion, Limitations and Directions for Future Work

Researchers have highlighted the relationship between data quality and decision support systems, and in particular recommender systems, in the field of IS. Based on a theoretical model, we present a procedure for the systematic extension of a data set DS1 with additional item content (attributes and attribute values) from another data set DS2 in the same domain. Thereby, the procedure aims to address data quality, especially by increasing the completeness of data sets and, in consequence, to improve recommendation quality of recommender systems. In a first step, an approach to detect duplicate items across data sets DS1 and DS2 is proposed. In a second step, we outline how item content data in DS1 can be extended by integrating the item content data of a data set DS2 as well as by imputing missing values. Based on these two steps, the resulting extended data set can be used by an arbitrary content-based or hybrid recommender system to determine recommendations in a subsequent step. We evaluate the procedure by using two real-world data sets regarding restaurants and movies, which constitute commonly analyzed domains in IS research on e-commerce, and discuss effects on recommendation quality. Here, the results show that the presented procedure is indeed capable of improving recommendations considerably by means of item content data extension, which is in line with existing research (cf. Heinrich et al. 2019). Furthermore, we investigate different effects on the results of the procedure from the three dimensions items, content and users, revealing that the procedure was valuable in each investigated case and indicating under which circumstances a substantial improvement in recommendation quality was achieved. Complementary to existing research proposing general relationships between data quality and decision support systems, this work provides and evaluates a tangible procedure which enables to increase data completeness with the aim of improving recommendation quality. Moreover, this procedure serves an evaluated template for future procedures to support the investigation of further data quality dimensions (e.g., consistency) for decision support systems in various e-commerce applications.

The rapid proliferation of e-commerce has cemented the tremendous relevance of recommender systems. These systems are powerful data-driven decision support systems incorporated in many e-commerce platforms guiding users to their individually best item choice among a plethora of alternatives. Thereby, recommender systems address the problem of information overload, which constitutes a major subject of IS research in the field of e-commerce. While the steady increasing volume of information (e.g., about attributes of items) would further aggravate the problem of information overload for users, recommender systems actually can somehow invert this effect. In contrast to the limited cognitive capabilities of users, for recommender systems as automated data-driven systems, more information (e.g., item content data; i.e., attributes and attribute values) is highly useful to individually support the user's decision-making and thus to further reduce the problem of information overload. To do so, increasing the completeness of the data (i.e., item content data) a recommender system is based on seems to constitute a promising way, which is studied in this paper by proposing a procedure for data set extension. Especially in established e-commerce domains (e.g., restaurants and movies), a higher completeness can significantly improve the recommendation quality for users (e.g., the selection of restaurants and movies), which in the long run strengthens the relationship between providers and users.

Here, our evaluation encourages IS providers in e-commerce (e.g., online portals) to improve data quality by providing a straightforward way to increase completeness without the need of manual tasks such as visiting items' websites or social media pages. Our procedure shows that achieving high data quality is indeed beneficial for companies, as the resulting improved recommendations support the various goals and purposes of recommender systems such as promoting cross- and up-selling or increasing customer loyalty (Jannach and Adomavicius 2016). Moreover, our results open up a way for portals with limited resources to balance the efforts and benefits associated to the procedure. For instance, as recommending items based on massively extended item content data can prove to be time-consuming, portals may prefer to focus on a subset of users or additional attributes based on the evidence found in Section 4.

However, our work also has some limitations, which could be starting points for future research. First, while we focused on completeness as a highly relevant data quality dimension, extensions of data sets in the context of recommender systems could also take into account other data quality dimensions such as accuracy or currency. Second, we considered the extension of item content data based on additional structured data in this paper. Here, it would be promising to leverage modern information extraction approaches, such as aspect extraction with language models (e.g., BERT; cf. Xu et al. 2019). Thereby, data sets already used by IS providers could be extended by extracted features from unstructured textual data sources (e.g., online customer reviews). Moreover, another interesting perspective might be to incorporate the extension of user data into the procedure, which could in some cases be realized by, for instance, user linkage based on online social network accounts. Finally, the approach could also be applied to further data sets, possibly from other domains outside the field of e-commerce, in order to validate and substantiate the resulting effects on recommendation quality.

References

Abel, F., Herder, E., Houben, G.-J., Henze, N., & Krause, D. (2013). Cross-system user modeling and personalization on the Social Web. *User Modeling and User-Adapted Interaction, 23*, 169–209. https://doi.org/10.1007/s11257-012-9131-2.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*, 734–749. https://doi.org/10.1109/TKDE.2005.99.

Aggarwal, C. C. (2016). *Recommender Systems*. Cham: Springer International Publishing.

Amatriain, X., Pujol, J. M., Tintarev, N., & Oliver, N. (2009). Rate it again. In L. Bergman, A. Tuzhilin, R. Burke, A. Felfernig, & L. Schmidt-Thieme (Eds.), *The third ACM conference on Recommender systems, New York, New York, USA* (pp. 173–180). New York, NY: ACM. https://doi.org/10.1145/1639714.1639744.

Basaran, D., Ntoutsi, E., & Zimek, A. (2017). Redundancies in Data and their Effect on the Evaluation of Recommendation Systems: A Case Study on the Amazon Reviews Datasets. In N. Chawla & W. Wang (Eds.), *The 2017 SIAM International Conference on Data Mining, Houston, Texas, USA* (pp. 390–398). Philadelphia, PA: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974973.44.

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality*. Cham: Springer International Publishing.

Berkovsky, S., Kuflik, T., & Ricci, F. (2012). The impact of data obfuscation on the accuracy of collaborative filtering. *Expert Systems with Applications, 39*, 5033–5042. https://doi.org/10.1016/j.eswa.2011.11.037.

Bharati, P., & Chaudhury, A. (2004). An empirical investigation of decision-making satisfaction in web-based decision support systems. *Decision Support Systems, 37*, 187–197. https://doi.org/10.1016/S0167-9236(03)00006-X.

Blake, R., & Mangiameli, P. (2011). The Effects and Interactions of Data Quality and Problem Complexity on Classification. *Journal of Data and Information Quality, 2*, 1–28. https://doi.org/10.1145/1891879.1891881.

Bleiholder, J., & Naumann, F. (2008). Data fusion. *ACM Computing Surveys, 41*, 1–41. https://doi.org/10.1145/1456650.1456651.

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012). TasteWeights: a visual interactive hybrid recommender system. In P. Cunningham, N. Hurley, I. Guy, & S. S. Anand (Eds.), *The sixth ACM conference on Recommender systems, Dublin, Ireland* (pp. 35–42). New York, NY: ACM. https://doi.org/10.1145/2365952.2365964.

Bouadjenek, M. R., Pacitti, E., Servajean, M., Masseglia, F., & Abbadi, A. E. (2018). A Distributed Collaborative Filtering Algorithm Using Multiple Data Sources. *arXiv preprint arXiv:1807.05853*.

Bunnell, L., Osei-Bryson, K.-M., & Yoon, V. Y. (2019). RecSys Issues Ontology: A Knowledge Classification of Issues for Recommender Systems Researchers. *Information Systems Frontiers, 97*, 667. https://doi.org/10.1007/s10796-019-09935-9.

Burke, R., & Ramezani, M. (2011). Matching Recommendation Technologies and Domains. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 367–386). Boston, MA: Springer US.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28.

Chang, J.-H., Tsai, C.-E., & Chiang, J.-H. (2018). Using Heterogeneous Social Media as Auxiliary Information to Improve Hotel Recommendation Performance. *IEEE Access, 6*, 42647–42660. https://doi.org/10.1109/ACCESS.2018.2855690.

Chang, W.-L., & Jung, C.-F. (2017). A hybrid approach for personalized service staff recommendation. *Information Systems Frontiers, 19*, 149–163. https://doi.org/10.1007/s10796-015-9597-7.

Christen, P. (2012). *Data matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management, 20*, 17–28. https://doi.org/10.1016/S0268-4012(99)00051-1.

Enders, C. K. (2010). *Applied missing data analysis* (Methodology in the social sciences). New York: Guilford Press.

Feldman, M., Even, A., & Parmet, Y. (2018). A methodology for quantifying the effect of missing data on decision quality in classification problems. *Communications in Statistics–Theory and Methods, 47*(11), 2643–2663.

Forbes, P., & Zhu, M. (2011). Content-boosted matrix factorization for recommender systems. In B. Mobasher, R. Burke, D. Jannach, & G. Adomavicius (Eds.), *The fifth ACM conference on Recommender systems, Chicago, Illinois, USA* (pp. 261–264). New York, NY: ACM. https://doi.org/10.1145/2043932.2043979.

Ge, M. (2009). *Information quality assessment and effects on inventory decision-making*. Doctoral dissertation. Dublin City University, Dublin.

GitHub. (2020). Procedure Completeness: Extending Item Content Data. https://github.com/ProcedureCompleteness/ExtendingItemContentDataSets. Accessed 14 September 2020.

Hasan, M. R., Jha, A. K., & Liu, Y. (2018). Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives. *Computers in Human Behavior, 80*, 220–228. https://doi.org/10.1016/j.chb.2017.11.020.

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., & Szubartowicz, M. (2019). Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems. *Electronic Markets, 23*, 169. https://doi.org/10.1007/s12525-019-00366-7.

Heinrich, B., Klier, M., Schiller, A., & Wagner, G. (2018). Assessing data quality – A probability-based metric for semantic consistency. *Decision Support Systems, 110*, 95–106. https://doi.org/10.1016/j.dss.2018.03.011.

Jannach, D., & Adomavicius, G. (2016). Recommendations with a Purpose. In S. Sen & W. Geyer (Eds.), *The 10th ACM Conference on Recommender Systems, Boston, Massachusetts, USA* (pp. 7–10). New York, NY, USA: Association for Computing Machinery.

Jannach, D., Zanker, M., Ge, M., & Gröning, M. (2012). Recommender Systems in Computer Science and Information Systems – A Landscape of Research. *E-Commerce and Web Technologies, 123*, 76–87. https://doi.org/10.1007/978-3-642-32273-0_7.

Jurek, A., Hong, J., Chi, Y., & Liu, W. (2017). A novel ensemble learning approach to unsupervised record linkage. *Information Systems, 71*, 40–54. https://doi.org/10.1016/j.is.2017.06.006.

Kamath, K. Y., Caverlee, J., Lee, K., & Cheng, Z. (2013). Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In D. Schwabe (Ed.), *The 22nd International Conference on the World Wide Web, Rio de Janeiro, Brazil* (pp. 667–678). New York, NY: ACM. https://doi.org/10.1145/2488388.2488447.

Kamis, A., Stern, T., & Ladik, D. M. (2010). A flow-based model of web site intentions when users customize products in business-to-consumer electronic commerce. *Information Systems Frontiers, 12*, 157–168. https://doi.org/10.1007/s10796-008-9135-y.

Karimova, F. (2016). A Survey of e-Commerce Recommender Systems. *European Scientific Journal, ESJ, 12*, 75. https://doi.org/10.19044/esj.2016.v12n34p75.

Karumur, R. P., Nguyen, T. T., & Konstan, J. A. (2018). Personality, User Preferences and Behavior in Recommender systems. *Information Systems Frontiers, 20*, 1241–1265. https://doi.org/10.1007/s10796-017-9800-0.

Kayaalp, M., Özyer, T., & Özyer, S. T. (2009). A Collaborative and Content Based Event Recommendation System Integrated with Data Collection Scrapers and Services at a Social Networking Site. In N. Memon (Ed.), *International Conference on Advances in Social Networks Analysis and Mining, 2009, Athens, Greece* (pp. 113–118). Piscataway, NJ: IEEE. https://doi.org/10.1109/ASONAM.2009.41.

Kim, D., Park, C., Oh, J., Lee, S., & Yu, H. (2016). Convolutional Matrix Factorization for Document Context-Aware Recommendation. In S. Sen, W. Geyer, J. Freyne, & P. Castells (Eds.), *The 10th ACM Conference on Recommender Systems, Boston, Massachusetts, USA* (pp. 233–240). New York, New York, USA: ACM Press. https://doi.org/10.1145/2959100.2959165.

Koren, Y. (2009). The bellkor solution to the netflix grand prize. *Netflix prize documentation, 81*, 1–10.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer, 42*, 30–37. https://doi.org/10.1109/MC.2009.263.

Lathia, N., Amatriain, X., & Pujol, J. M. (2009). Collaborative filtering with adaptive information sources. In S. S. Anand, B. Mobasher, A. Kobsa, & D. Jannach (Eds.), *7th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems, Pasadena, California, USA* (pp. 81–86, CEUR Workshop Proceedings (CEUR-WS.org), Vol. 528).

Levi, A., Mokryn, O., Diot, C., & Taft, N. (2012). Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In P. Cunningham, N. Hurley, I. Guy, & S. S. Anand (Eds.), *The sixth ACM conference on Recommender systems, Dublin, Ireland* (pp. 115–122). New York, NY: ACM. https://doi.org/10.1145/2365952.2365977.

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science, 9*, 181–212.

Li, Y., Zhang, Z., Peng, Y., Yin, H., & Xu, Q. (2018). Matching user accounts based on user generated content across social networks. *Future Generation Computer Systems, 83*, 104–115. https://doi.org/10.1016/j.future.2018.01.041.

Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems, 74*, 12–32. https://doi.org/10.1016/j.dss.2015.03.008.

Manca, M., Boratto, L., & Carta, S. (2018). Behavioral data mining to produce novel and serendipitous friend recommendations in a social bookmarking system. *Information Systems Frontiers, 20*, 825–839. https://doi.org/10.1007/s10796-015-9600-3.

Mladenić, D., & Grobelnik, M. (2003). Feature selection on hierarchy of web documents. *Decision Support Systems, 35*, 45–87. https://doi.org/10.1016/S0167-9236(02)00097-0.

Molina, L. C., Belanche, L., & Nebot, À. (2002). Feature selection algorithms: a survey and experimental evaluation. In V. Kumar (Ed.), *IEEE International Conference on Data Mining, Maebashi City, Japan* (pp. 306–313). Los Alamitos, CA: IEEE Computer Society.

Naumann, F., Freytag, J.-C., & Leser, U. (2004). Completeness of integrated information sources. *Information Systems, 29*, 583–615. https://doi.org/10.1016/j.is.2003.12.005.

Nguyen, J., & Zhu, M. (2013). Content-boosted matrix factorization techniques for recommender systems. *Statistical Analysis and Data Mining, 6*, 286–301. https://doi.org/10.1002/sam.11184.

Nguyen, T. T., Maxwell Harper, F., Terveen, L., & Konstan, J. A. (2018). User Personality and User Satisfaction with Recommender Systems. *Information Systems Frontiers, 20*, 1173–1189. https://doi.org/10.1007/s10796-017-9782-y.

Ning, Y., Shi, Y., Hong, L., Rangwala, H., & Ramakrishnan, N. (2017). A Gradient-based Adaptive Learning Framework for Efficient Personal Recommendation. In P. Cremonesi, F. Ricci, S. Berkovsky, & A. Tuzhilin (Eds.), *The Eleventh ACM Conference on Recommender Systems, Como, Italy* (pp. 23–31). New York, New York, USA: ACM Press. https://doi.org/10.1145/3109859.3109909.

Ntoutsi, E., & Stefanidis, K. (2016). Recommendations beyond the ratings matrix. In Association for Computing Machinery (Ed.), *The Workshop on Data-Driven Innovation on the Web, Hannover, Germany* (pp. 1–5). New York, New York, USA: ACM Press. https://doi.org/10.1145/2911187.2914580.

Ozsoy, M. G., Polat, F., & Alhajj, R. (2016). Making recommendations by integrating information from multiple social networks. *Applied Intelligence, 45*, 1047–1065. https://doi.org/10.1007/s10489-016-0803-1.

Peska, L., & Vojtas, P. (2015). Using Implicit Preference Relations to Improve Content Based Recommending. *E-Commerce and Web Technologies, 239*, 3–16. https://doi.org/10.1007/978-3-319-27729-5_1.

Pessemier, T. de, Dooms, S., Deryckere, T., & Martens, L. (2010). Time dependency of data quality for collaborative filtering algorithms. In X. Amatriain, M. Torrens, P. Resnick, & M. Zanker (Eds.), *The fourth ACM conference on Recommender systems, Barcelona, Spain* (pp. 281–284). New York, NY: ACM. https://doi.org/10.1145/1864708.1864767.

Picault, J., Ribiere, M., Bonnefoy, D., & Mercer, K. (2011). How to get the Recommender out of the Lab? In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 333–365). Boston, MA: Springer US.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*, 211–218. https://doi.org/10.1145/505248.506010.

Porcel, C., & Herrera-Viedma, E. (2010). Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries. *Knowledge-Based Systems, 23*(1), 32–39.

Power, D. J., Sharda, R., & Burstein, F. (2015). *Decision support systems*. Hoboken, New Jersey, USA: John Wiley & Sons, Ltd.

Raad, E., Chbeir, R., & Dipanda, A. (2010). User Profile Matching in Social Networks. In T. Enokido (Ed.), *13th International Conference on Network-Based Information Systems (NBIS), 2010* (pp. 297–304). Piscataway, NJ: IEEE Service Center.

Ricci, F., Rokach, L., & Shapira, B. (Eds.). (2015a). *Recommender Systems Handbook*. Boston, MA: Springer US.

Ricci, F., Rokach, L., & Shapira, B. (2015b). Recommender Systems: Introduction and Challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 1–34). Boston, MA: Springer US.

Richthammer, C., & Pernul, G. (2018). Situation awareness for recommender systems. *Electronic Commerce Research, 37*, 85. https://doi.org/10.1007/s10660-018-9321-z.

Sar Shalom, O., Berkovsky, S., Ronen, R., Ziklik, E., & Amihood, A. (2015). Data Quality Matters in Recommender Systems. In H. Werthner, M. Zanker, J. Golbeck, & G. Semeraro (Eds.), *9th ACM Conference on Recommender Systems, Vienna, Austria* (pp. 257–260). New York, NY: ACM. https://doi.org/10.1145/2792838.2799670.

Scannapieco, M., & Batini, C. (2004). Completeness in the Relational Model: a Comprehensive Framework. In *International Conference on Information Quality, Cambridge, Massachusetts, USA* (pp. 333–345).

Scholz, M., Dorner, V., Schryen, G., & Benlian, A. (2017). A configuration-based recommender system for supporting e-commerce decisions. *European Journal of Operational Research, 259*(1), 205–215.

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 257–297). Boston, MA: Springer US.

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon. com. *Ieee internet computing, 21*(3), 12–18.

Statista. (2019). Statistics and Market Data about E-commerce. https://www.statista.com/markets/413/e-commerce/. Accessed 3 June 2020.

Steorts, R. C., Ventura, S. L., Sadinle, M., & Fienberg, S. E. (2014). A Comparison of Blocking Methods for Record Linkage. In J. Domingo-Ferrer (Ed.), *Privacy in Statistical Databases* (Vol. 8744, pp. 253–268, Lecture Notes in Computer Science). Cham: Springer International Publishing.

Tang, H., Lee, C. B. P., & Choong, K. K. (2017). Consumer decision support systems for novice buyers – a design science approach. *Information Systems Frontiers, 19*, 881–897. https://doi.org/10.1007/s10796-016-9639-9.

Vanaja, R., & Mukherjee, S. (2019). Novel Wrapper-Based Feature Selection for Efficient Clinical Decision Support System. In L. Akoglu, E. Ferrara, M. Deivamani, R. Baeza-Yates, & P. Yogesh (Eds.), *Third International Conference on Intelligent Information Technologies, Chennai, India* (Vol. 941, pp. 113–129,

Communications in Computer and Information Science, Vol. 941). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-3582-2_9.

Vargas-Govea, B., González-Serna, G., & Ponce-Medellın, R. (2011). Effects of relevant contextual features in the performance of a restaurant recommender system. In B. Mobasher, R. Burke, D. Jannach, & G. Adomavicius (Eds.), *The fifth ACM conference on Recommender systems, Chicago, Illinois, USA* (pp. 592–596). New York, NY: ACM.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*, 86–95. https://doi.org/10.1145/240455.240479.

Wei, C., Khoury, R., & Fong, S. (2013). Web 2.0 Recommendation service by multi-collaborative filtering trust network algorithm. *Information Systems Frontiers, 15*, 533–551. https://doi.org/10.1007/s10796-012-9377-6.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods, Alexandria, Virginia*. Alexandria, Virginia, USA: American Statistical Association.

Woodall, P., Borek, A., Gao, J., Oberhofer, M., & Koronios, A. (2015). An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics. In R. Wang (Ed.), *19th International Conference on Information Quality, Xi'an, China* (pp. 24–33, Management and data quality). Red Hook, NY: Curran.

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In (pp. 2324–2335). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1242.

Yan, X., Wang, J., & Chau, M. (2015). Customer revisit intention to restaurants: Evidence from online reviews. *Information Systems Frontiers, 17*, 645–657. https://doi.org/10.1007/s10796-013-9446-5.

Zhou, L. (2020). Product advertising recommendation in e-commerce based on deep learning and distributed expression. *Electronic Commerce Research, 20*, 321–342. https://doi.org/10.1007/s10660-020-09411-6.

3.2 Paper: Leveraging Fine-grained Supervision to Improve Multiple Instance Learning for Fine-grained Sentiment Classification in Online Consumer Reviews

Current Status	**Citation**
This paper was **under review** at *30th European Conference on Information Systems*. After rework, it will be submitted to *Transactions of the Association for Computational Linguistics.*	Hopf, M. (2022). "Leveraging Fine-grained Supervision to Improve Multiple Instance Learning for Fine-grained Sentiment Classification in Online Consumer Reviews" Working Paper, University of Regensburg.

LEVERAGING FINE-GRAINED SUPERVISION TO IMPROVE MULTIPLE INSTANCE LEARNING FOR FINE-GRAINED SENTIMENT CLASSIFICATION IN ONLINE CONSUMER REVIEWS

Research Paper

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

Abstract

High volumes of consumer reviews are accessible, which constitute important performance indicators for businesses and impactful purchase factors for consumers, as reviews comprise fine-grained textual consumer sentiments regarding items as well as an associated star rating. For enabling fine-grained sentiment classification for high volumes of consumer reviews in an economical manner, extant literature proposes multiple instance learning (MIL). The basic idea of MIL is to leverage the relationships between review texts and associated star ratings. Here, MIL only uses review-level supervision for learning to predict fine-grained sentiments. Hence, it is promising that extending MIL with partly supervision on few labeled data instances could improve the performance of MIL. In this paper, we propose MILPS, which extends MIL with fine-grained supervision, and analyze whether MILPS can improve performance for fine-grained sentiment classification. In our evaluation, we find that MILPS indeed enables improved performance compared to MIL, while preserving its economical benefits.

1 Introduction

Online consumer reviews constitute important performance indicators for businesses, impactful factors for consumer purchase decisions as well as vital information for review platforms to infer user preferences (Siering et al., 2018). Huge amounts of these reviews are generated on a daily basis from millions of users regarding various items (e.g., products or services) on several channels, such as review platforms and other social media channels. Further, such reviews are also subject of many research fields such as consumer behavior analysis or design of electronic-word-of-mouth (EWOM) systems (Jabr et al., 2020). Online consumer reviews are highly relevant for various stakeholders as they comprise a consumer generated review text, containing fine-grained sentiments of consumers regarding items (e.g., the sentences "Food is good. Drinks are expensive." in a restaurant review), and an associated star rating, representing the overall assessment of consumers (e.g., 4 stars). In particular, the review texts are highly valuable for businesses as well as for review platforms, as they constitute direct consumer feedback at high volume and enable to derive insights regarding consumers and items. However, because of the large volume of online consumer reviews available, a manual analysis of review texts regarding the contained sentiments is not feasible calling for an automated analysis. This automated analysis and classification of consumer sentiments in texts by means of algorithms is the main goal of the (aspect-based) sentiment analysis (Pontiki et al., 2016), a subfield of research regarding natural language processing (NLP). The standard technique for fine-grained sentiment classification is supervised classification by means of training an NLP model (e.g., the language model BERT by Devlin et al., 2019) on review text segments, which are labeled with a sentiment, and utilizing this model for the prediction of sentiments of new text segments (Pannala et al., 2016). Here, the performance of this supervised technique highly depends on the amount of labeled training data, which requires time-consuming human efforts for reading and annotating review text segments with sentiment labels. To mitigate this problem of highly time-consuming annotations for large amount of training data to achieve good performances for fine-grained sentiment classification, novel techniques based on multiple instance learning (MIL) have been developed for economical fine-grained sentiment classification in online consumer reviews (Angelidis and Lapata, 2018; Kotzias et al., 2015; Pappas and Popescu-Belis, 2017). The basic idea of MIL is to infer fine-grained sentiments for review text segments by using the relationship between the review text and the associated star rating. In particular, it is assumed that the sentiments of multiple instances (i.e., the segments of the review text) are directly related to the rating, constituting the

sentiment label for the group of instances. Using MIL for predicting sentiments for the instances, MIL can be viewed as "backpropagating" the coarse-grained review-level sentiment information (i.e., the star rating) onto the fine-grained sentiments of review text segments. This means that MIL learns to predict fine-grained sentiments with a review-level supervision (Angelidis and Lapata, 2018). Because of the high volume of review data available on review platforms, MIL is well suited for economical fine-grained sentiment classification in online consumer reviews. The advantage of MIL models compared to the supervised technique is that no costly labeling efforts are required at all.

Nevertheless, using supervised fine-grained sentiment classification models trained on data with fine-grained sentiment labels is very precise and well established in the literature (Pappas and Popescu-Belis, 2017) due to very good performances (Hoang et al., 2019). Given these complementary advantages of models with fine-grained supervision and MIL models with review-level supervision, in this paper we investigate how to combine both techniques to a fine-grained sentiment classification technique based on MIL with partly fine-grained supervision (MILPS) and whether this MILPS can improve the performance of MIL. The idea of our approach is similar to semi-supervised learning, which combines few labeled data as well as large amounts of unlabeled data to perform certain machine learning tasks (van Engelen and Hoos, 2020). Thereby, semi-supervised learning allows to utilize the large amounts of available unlabeled data available in combination with smaller sets of labeled data. In contrast to semi-supervised learning, the envisioned approach in this paper utilizes large amounts of review-level labeled data (i.e., reviews with associated star ratings). Nevertheless, our proposed approach is especially useful in scenarios with datasets that only contain few labeled data (i.e., fine-grained sentiment labels for text instances). In this paper, we therefore investigate the following research questions:

> **RQ1.** How can instance labels be incorporated into MIL models to obtain MILPS models for fine-grained sentiment classification?
>
> **RQ2.** Does instance label data in MILPS models improve the performance of fine-grained sentiment classification compared to MIL models and supervised models?

To address these research questions, we propose a MILPS model that extends MIL with partly fine-grained supervision by incorporating instance sentiment labels, and analyze and compare the performance of a state-of-the-art supervised model, a state-of-the-art MIL model and the MILPS model on a dataset comprising online consumer reviews for restaurants. By our evaluation, we find that the MILPS model enables improved performance of fine-grained sentiment classification compared to the MIL model with highly reduced labeling efforts compared to the supervised model, which is therefore cost-efficient and economical.

The remainder of the paper is structured as follows. In the next section, we outline the background for fine-grained sentiment classification regarding supervised and MIL models as well as discuss the related work. Then, we outline the proposed MILPS model, which shall enable economical fine-grained sentiment classification as achieved by MIL models, but with higher classification performance as achieved by supervised models. In the fourth section, we outline how we evaluated the proposed MILPS model on a real-world dataset of restaurant reviews and state the evaluation results. Thereafter, we critically discuss the evaluation results, state the major findings and derive managerial implications of our research. In the last section, we conclude by summarizing the conducted research and depicting limitations that could be starting points for future works.

2 Background and Related Work

In this section, we recall the basic knowledge and notation for fine-grained sentiment classification and outline how fine-grained sentiment classification can be executed with supervised models and with MIL models. After that, we discuss the related work and identify the research gap addressed by this paper.

2.1 Fine-grained sentiment classification

Reviews can be divided into fine-grained instances x_i (i.e., disjunct text segments), depending on a segmentation policy (e.g., each sentence is an instance) and it can be assumed that each instance has a sentiment. The goal of fine-grained sentiment classification is to predict a sentiment class label $p_i \in \{-1,0,1\}$, for each instance x_i, with $-1,0,1$ representing a *negative*, *neutral*, *positive* sentiment, respectively.

2.2 Supervised fine-grained sentiment classification

To predict sentiment classes p_i for instances x_i in a supervised manner, a model M_{sup} with parameters θ_{sup} is trained on a labeled dataset $d_{labeled}$ (i.e., $p_i = M_{sup}(x_i; \theta_{sup})$). The labeled dataset $d_{labeled}$ comprises tuples (x_i, s_i) of instances x_i with its true sentiment labels $s_i \in \{-1,0,1\}$ that are given by human annotators. In the training phase, a loss function $\mathcal{L}_{super}(d_{labeled}, M_{sup}, \theta_{sup})$ is minimized to learn the fine-grained sentiment classification model M_{sup}. Here, the loss function depends on the labeled dataset for training, the supervised model and its parameters. By minimizing the loss function, it is aimed to learn the parameters θ_{sup} such that the predicted sentiments approximate the true sentiments (i.e., $s_i \approx M_{sup}(x_i; \theta_{sup}) = p_i$) for instances x_i of $d_{labeled}$. Then, a prediction p_i for a new instance x_i is made by $p_i = M_{sup}(x_i; \theta_{sup})$.

2.3 Fine-grained sentiment classification with multiple instance learning

In contrast to supervised fine-grained sentiment classification that uses instance-level sentiment labels, MIL takes advantage of the high volume of coarse-grained, review-level sentiment labels (i.e., star ratings of reviews). The basic idea of MIL is to infer instance-level sentiments by using their relationship to the review-level sentiment. In particular, MIL can be viewed as "backpropagating" the coarse-grained sentiment information on the review-level to the fine-grained sentiments on the instance-level. Following works such as Angelidis and Lapata (2018), we outline the framework of MIL in more detail in the subsequent paragraph.

MIL is based on a dataset d_{MIL} consisting of reviews (rev_k, r_k) that each comprise a consumer generated review text rev_k and an associated discrete star rating $r_k \in \{1,2,3,4,5\}$, representing the overall sentiment of this review (i.e., the review-level sentiment). In particular, a review text is a sequence of instances $(x_{k1}, \dots, x_{kn_k}) = rev_k$ (e.g., sentences, depending on the segmentation policy; cf. Section 2.1). For ease of notation, we omit the index k and write x_i instead of x_{ki}. Here, it is important to note that these instances x_i do not have any sentiment class label in the dataset d_{MIL}. However, all of these instances x_i have a latent sentiment that can be associated to the classes negative, neutral or positive, hence for each x_i a sentiment class label $M_{MIL}(x_i; \theta_{MIL}) = p_i \in \{-1,0,1\}$ can be predicted by a MIL sentiment classification model M_{MIL}. Since the sentiment class labels are not given in d_{MIL}, the parameters θ_{MIL} of the MIL model cannot be estimated properly solely based on the instance data in d_{MIL}. Here, the advantage of MIL comes into play. MIL is based on the assumption that there exists an aggregation function f with parameters θ_f, such that $f(p_1, \dots, p_n; \theta_f) \approx r_k$, which means that the overall sentiment of a review can be computed by the fine-grained sentiments p_i of the instances x_i via a function f. Since the labels of the instances are not given in the dataset d_{MIL}, the function and the given ratings r_k are used to compute predictions for the sentiment classes p_i of instances x_i, in the sense of propagating the review-level sentiment information back to the instance level. Hence, in the training phase, a loss function $\mathcal{L}_{MIL}(d_{MIL}, M_{MIL}, \theta_{MIL}, f, \theta_f)$ is minimized to learn the sentiment classification model M_{MIL} and the aggregation function f. Here, the loss function depends on the dataset with unlabeled instances of reviews and the corresponding review-level star ratings for training, the MIL sentiment classification model, the aggregation function and their parameters. By minimizing the loss function, it is aimed to learn the parameters θ_{MIL} and θ_f such that the predicted review-level sentiment (i.e., the predicted sentiments aggregated by f) approximate the true review rating (i.e., $r_k \approx f(p_{k1}, \dots, p_{kn_k}; \theta_f) = f(M_{MIL}(x_{k1}; \theta_{MIL}), \dots, M_{MIL}(x_{kn_k}; \theta_{MIL}); \theta_f)$) for reviews $rev_k = (x_{k1}, \dots, x_{kn_k})$ of d_{MIL}. Then, in the prediction phase, a sentiment prediction p_i for a new instance x_i is made by $p_i = M_{MIL}(x_i; \theta_{MIL})$.

2.4 Related work

To prepare the related work, we followed the standard approach for creating a literature review as proposed by Levy and Ellis (2006). We searched the databases ACM Digital Library, AIS Library, IEEE Xplore, ScienceDirect, Springer Link and Wiley and used the search engine Google Scholar for the search term *("fine-grained" OR "sentence-level" OR "aspect-based") ("sentiment classification" OR "opinion classification") "rating" "multiple instance learning" "semi-supervised"*. This search led to 29 papers, which were manually screened based on title and abstract. Where necessary, we conducted a more detailed

analysis to assess the papers' relevancy. Based on the relevant papers, we conducted a further forward and backward search. In total, we identified 6 papers as relevant for our research.

All of these 6 papers propose MIL approaches for sentiment classification in online consumer reviews. While the works of Angelidis and Lapata (2018), Kotzias et al. (2015), Lutz et al. (2019), Pappas and Popescu-Belis (2017) and Pappas and Popescu-Belis (2014) aim at fine-grained sentiment classification of review text segments (e.g., sentences, elementary discourse units or aspect phrases) by means of MIL with review ratings, Correia et al. (2016) aim at sentiment classification of the whole review text by means of MIL with overall item ratings. Therefore, the approach of Correia et al. (2016) cannot be applied for fine-grained sentiment classification, as aimed for in the paper at hand. Further, Kotzias et al. (2015) and Lutz et al. (2019) propose an unweighted average aggregation function for MIL, which means that the review-level sentiment is the average of the classified fine-grained sentiments in the review text. In contrast, Angelidis and Lapata (2018), Pappas and Popescu-Belis (2017) and Pappas and Popescu-Belis (2014) use a weighted average of fine-grained sentiment classifications as aggregation function to obtain a review-level sentiment. Since it is reasonable that not all fine-grained sentiments in a review have the same impact on the review-level sentiment, weighted averages are better suited for fine-grained sentiment classification, which is supported by superior performance of the weighted average in the evaluation of Angelidis and Lapata (2018). Therefore, and since the MIL approach of Angelidis and Lapata (2018) extends the works from Pappas and Popescu-Belis (2017) and Pappas and Popescu-Belis (2014), the MIL model of Angelidis and Lapata (2018) constitutes the state-of-the-art for fine-grained sentiment classification with MIL. However, while all these works show promising results of MIL for sentiment classification, none of them considers supervision on the fine-grained level (e.g., sentiments for individual sentences). Thus, these works leave out vital potential of fine-grained sentiment information which seems promising for further enhancing the performance of MIL for fine-grained sentiment classification.

3 Multiple Instance Learning with Partly Fine-Grained Supervision for Fine-Grained Sentiment Classification

In this section, we outline the proposed MILPS approach for fine-grained sentiment classification. First, we start by introducing the basic idea that enables to use fine-grained supervision for MIL. After that, we specify and reason our chose for the precise MILPS model proposed by this paper, which will be analyzed in the evaluation section.

3.1 Basic idea of MILPS for fine-grained sentiment classification

To leverage fine-grained supervision for MIL, we need a dataset d_{MILPS} that consists of an integrated combination of datasets $d_{labeled}$ and d_{MIL} from Section 2.2 and Section 2.3 with overlapping instances x_i of both datasets. More precisely, d_{MILPS} comprises a (large) dataset d_{MIL} and a (small) dataset $d_{labeled}$, such that all labeled instances x_i of $d_{labeled}$ (i.e., $(x_i, s_i) \in d_{labeled}$) are contained in a review rev_k (i.e., $x_i \in rev_k = (x_{k1}, \dots x_{kn_k})$) of d_{MIL}. Hence and in contrast to d_{MIL}, there are instances x_i in d_{MILPS} that have the human annotated fine-grained sentiment class labels s_i. Thus, a sentiment class p_i for an instance x_i can be predicted by a partly supervised model $M_{ptly\ sup}$ with parameters θ_{sup} that is trained on $d_{labeled}$ (i.e., $p_i = M_{ptly\ sup}(x_i; \theta_{ptly\ sup})$), whereby $d_{labeled}$ is in general a small dataset, since labeling instances by human annotators is costly and time-consuming. However, each review from an online web portal has an associated star rating (i.e., a review-level sentiment), thus the size of d_{MIL} can be very large. In addition, it can be assumed that – similar to MIL (cf. Section 2.3) – there exists an aggregation function f with parameters θ_f, such that $f(p_1, \dots, p_n; \theta_f) \approx r_k$ for all $(rev_k, r_k) \in d_{MIL}$, which means that the overall sentiment r_k of a review $rev_k = (x_1, \dots, x_n)$ can be computed by the fine-grained sentiments p_i of the instances x_i via a function f. Hence, this function f and the given ratings r_k in d_{MIL} can also enable to compute predictions for the sentiment classes p_i of instances x_i and therefore, to train the parameters $\theta_{ptly\ sup}$ of the model $M_{ptly\ sup}$. Since the parameters $\theta_{ptly\ sup}$ of the model $M_{ptly\ sup}$ can be trained in a supervised manner on $d_{labeled}$, but also via MIL on d_{MIL}, without supervision on the instance level, we call this model *partly supervised*. Overall, to leverage both (i) the supervision on a small dataset $d_{labeled}$ of instances with sentiment labels and (ii) the high volume of review-level sentiment information in d_{MIL} for MIL, a loss function $\mathcal{L}_{MILPS}(d_{MIL}, d_{labeled}, M_{ptly\ sup}, \theta_{ptly\ sup}, f, \theta_f)$ is required that combines both approaches and enables to simultaneously learn the parameters $\theta_{ptly\ sup}$ and θ_f. Thus, by minimizing the

loss function, it is aimed (i) to learn the parameters $\theta_{ptly\ sup}$ such that the predicted sentiments approximate the true sentiments (i.e., $s_i \approx M_{ptly\ sup}(x_i; \theta_{ptly\ sup}) = p_i$) for instances x_i of $d_{labeled}$ as well as (ii) to learn the parameters $\theta_{ptly\ sup}$ and θ_f such that the predicted review-level sentiment (i.e., the predicted sentiments aggregated by f) approximate the true review rating (i.e., $r_k \approx f(p_{k1}, \dots, p_{kn_k}; \theta_f) = f(M_{ptly\ sup}(x_{k1}; \theta_{ptly\ sup}), \dots, M_{MIL}(x_{kn_k}; \theta_{ptly\ sup}); \theta_f)$) for reviews $rev_k = (x_{k1}, \dots, x_{kn_k})$ of d_{MIL}. Then, in the prediction phase, a sentiment class prediction for a new instance x_i is made by $p_i = M_{ptly\ sup}(x_i)$.

3.2 MILPS with BERT and MILNET for fine-grained sentiment classification

In this subsection, we specify the models M_{sup}, M_{MIL} and $M_{ptly\ sup}$ and the aggregation function f that we use for evaluation in the next section. Here, it is noteworthy that it is possible to choose the same initial model for M_{sup}, M_{MIL} and $M_{ptly\ sup}$, such that they only differ by their parameters θ_{sup}, θ_{MIL} and $\theta_{ptly\ sup}$ after training. Therefore, we only describe the architecture of the model M_{sup} in the following.

Recently, language models such as BERT (Devlin et al., 2019) have established themselves as state-of-the-art techniques in many NLP tasks (e.g., sentiment classification). Thus, we choose BERT for the model M_{sup}. BERT is a deep neural network, which is trained on huge datasets to provide semantic embeddings for words and sentences. A further advantage of BERT is that it is context-dependent, which means that BERT can differentiate between different meanings of one single word depending on the other words in a sentence. For an input text, BERT provides embeddings for each token of the text as well as an embedding for the whole text. To use BERT for classifying fine-grained sentiments for instances (i.e., text segments), we use the BERT architecture with one embedding for the whole instance. Then, it is only necessary to put a single classification layer on top of BERT's instance embedding and to fine tune this classification layer together with BERT's semantic embeddings. For a given embedding size H, BERT yields an embedding $BERT(x_i) = h_i \in \mathbb{R}^H$ for an input instance x_i. Then, as proposed by Devlin et al. (2019) a log-softmax layer is used for classification yielding a three-dimensional logit vector $\text{logit}_i = \log(\text{softmax}(Wh_i)) \in \mathbb{R}^3$ for an instance x_i, with weights $W \in \mathbb{R}^{3\times H}$. Here, the three dimensions of logit_i correspond to the three sentiment classes $\{-1,0,1\}$. The sentiment class prediction p_i for an instance x_i is then given by

$$p_i = M_{sup}(x_i; \theta_{sup}) = \text{argmax}_{k\in\{-1,0,1\}}(\text{logit}_i) \in \{-1,0,1\}.$$

This means, that θ_{sup} comprises the parameters of the language model BERT as well as the weights W of the log-softmax classification layer (analogously for M_{MIL} with θ_{MIL} and $M_{ptly\ sup}$ with $\theta_{ptly\ sup}$).

Next, we specify the aggregation function f. As outlined in Section 2.4, the MIL approach MILNET by Angelidis and Lapata (2018) constitutes the state-of-the-art MIL approach for fine-grained sentiment classification. Therefore, we adopt the aggregation function of Angelidis and Lapata (2018), which is a weighted average with weights computed by attention neural networks. More precisely, the aggregation function f yields a review-level sentiment prediction $r_k^\star$ for a review $rev_k = (x_1, \dots, x_n)$ with $p_i = M_Y(x_i; \theta_Y)$, $Y \in \{MIL, ptly\ sup\}$ and is given by

$$r_k^\star = f(p_1, \dots, p_n; \theta_f) = \sum_i \gamma_i * p_i \in [-1,1],$$

with attention weights $\gamma_i \in [0,1]$ for each instance x_i with $\sum_i \gamma_i = 1$. To scale the review-level sentiment prediction $r_k^\star$ to the same interval as the true star ratings of reviews $r_k \in [1,5]$, we compute $\hat{r}_k = (r_k^\star * 2) + 3 \in [1,5]$.

As described by Angelidis and Lapata (2018), the attention weights γ_i are computed with a bidirectional gated recurrent unit $BiGRU(h_i) \in \mathbb{R}^H$ and a softmax-tanh layer with weights $W_{att.} \in \mathbb{R}^{1\times H}$, $b_{att.} \in \mathbb{R}$ as given by

$$\gamma_i = \text{softmax}(\tanh(W_{att.} BiGRU(h_i) + b_{att.})) \in [0,1].$$

In the following, we will outline the loss functions $\mathcal{L}_{sup}$, $\mathcal{L}_{MIL}$ and $\mathcal{L}_{MILPS}$ To account for the ordered scale of sentiment classes, we use the mean absolute error (MAE) loss function for $\mathcal{L}_{sup}$, given by

$$\mathcal{L}_{sup}(d_{labeled}, M_{sup}, \theta_{sup}) = \sum_{(x_i,s_i)\in d_{labeled}} |p_i - s_i|.$$

Similar, we use a MAE loss function for MIL, given by

$$\mathcal{L}_{MIL}(d_{MIL}, M_{MIL}, \theta_{MIL}, f, \theta_f) = \sum_{(rev_k, r_k) \in d_{MIL}} |\hat{r}_k - r_k|.$$

The goal in training for supervised fine-grained sentiment classification is to minimize this loss function $\mathcal{L}_{sup}(d_{labeled}, M_{sup}, \theta_{sup})$ by learning the parameters θ_{sup}, while the goal for MIL is to minimize the loss function $\mathcal{L}_{MIL}(d_{MIL}, M_{MIL}, \theta_{MIL}, f, \theta_f)$ by learning the parameters θ_{MIL} and θ_f.

For MILPS, we choose the state-of-the-art techniques BERT and MILNET and we propose to combine the advantages of supervised learning based on fine-grained sentiment class labels and MIL on a large dataset with review-level sentiment data. Therefore, we propose to use the supervised model $M_{ptly\,sup} = M_{sup}$ and the attention-based MILNET model with the aggregation function $f(p_1, \dots, p_n; \theta_f) = \sum_i \alpha_i * p_i$. Inspired by the work of Nayak et al. (2020), who consider the similar problem of multi-view learning, where two related classification tasks are merged into one loss function by means of semi-supervised learning techniques, we specify the loss function $\mathcal{L}_{MILPS}$ for MILPS such that it combines both loss functions $\mathcal{L}_{sup}$ and $\mathcal{L}_{MIL}$. Here, $\mathcal{L}_{sup}$ can only be assessed for instances x_i from the dataset $d_{labeled}$. So, we set $\mathcal{L}_{sup}$ to zero for all x_i that do not have a fine-grained sentiment class label (i.e., for all $i \notin \{i|(x_i, s_i) \in d_{labeled}\} = I_{labeled}$). The loss function for MILPS is then given by

$$\mathcal{L}_{MILPS}(d_{MIL}, d_{labeled}, M_{ptly\,sup}, \theta_{ptly\,sup}, f, \theta_f) =$$
$$= \alpha * \mathcal{L}_{MIL}(d_{MIL}, M_{ptly\,sup}, \theta_{ptly\,sup}, f, \theta_f) + (1-\alpha) * \mathcal{L}_{sup}(d_{labeled}, M_{ptly\,sup}, \theta_{ptly\,sup}) =$$
$$= \alpha * \sum_{(rev_k, r_k) \in d_{MIL}} |\hat{r}_k - r_k| + (1-\alpha) * \sum_{i \in I_{labeled}} |p_{i,superv.} - s_i|.$$

Here, $\alpha \in [0,1]$ is a weighting parameter that allows to shift the focus of the loss function between the MIL and supervised loss function. By means of this MILPS model, we answer the first research question.

4 Evaluation

In this section we outline the used dataset for evaluation and the specifications of training the models on the data. After that, we describe the used evaluation criteria and present the evaluation results.

4.1 Dataset

For evaluation, we used a real-world review dataset from the commonly utilized review domain of restaurants (cf. e.g., Angelidis and Lapata, 2018; Pontiki et al., 2015). Reviews of this domains are also used for analyses in related research fields such as consumer behavior analysis or design of EWOM systems (Gutt et al., 2019). This restaurant dataset consists of 1,266 reviews for restaurants, bars and cafés in New York City from an established web portal for reviews regarding local businesses. Further, each review (rev_k, r_k) consists of a textual consumer review rev_k with an associated star rating r_k on a five-tier scale from 1 star to 5 stars. Hence, the dataset of these reviews (rev_k, r_k) constitutes M_{MIL}. The textual reviews comprise in total 10,353 sentences and each sentence is assigned with a sentiment class label negative, neutral or positive. Therefore, we consider the sentences of reviews as instances x_i in our evaluation. The dataset further exhibits a "J-shaped" rating and fine-grained sentiment distribution as outlined in Table 1, which is typical for online consumer reviews (e.g., cf. Debortoli et al., 2016).

Reviews with …					Sentences with … Sentiment		
1 Star	**2 Stars**	**3 Stars**	**4 Stars**	**5 Stars**	**Negative**	**Neutral**	**Positive**
134	111	204	436	381	2,007	3,688	4,658

Table 1. Rating and fine-grained sentiment distribution of the dataset.

For evaluation of the models described in section 3.2, we divided the dataset by a 70%/15%/15% split resulting in 886 reviews (with 7,369 sentences) for training, 190 reviews (with 1,424 sentences) for validation of hyper parameters and 190 reviews (with 1,560 sentences) for testing the models. As in the train dataset, all instances do have a sentiment class label, we can vary the size of $d_{labeled}$ in d_{MILPS} and evaluate the performance depending on the amount of fine-grained instance labels. Therefore, we analyze different ratios β of instances that have a fine-grained sentiment class label in the data. Let $I_{labeled} =$

$\{x_i|(x_i, s_i) \in d_{labeled}\}$ be the set of instances with an associated sentiment class label in $d_{labeled}$. Further, let $I_{all} = \{x_i|\exists(rev_k, r_k) \in d_{MIL}\ s.t.\ x_i \in rev_k = (x_{k1}, \dots, x_{kn_k})\}$ be the set of all instances in d_{MIL}. Then, the ratio of labeled instances is given by $\beta = \frac{|I_{labeled}|}{|I_{all}|}$. To evaluate different values of β, we randomly sampled reviews (rev_k, r_k) from d_{MIL} and put them into $d_{labeled,\beta}$ until β reaches the desired value (e.g., $\beta = 1\%$). To ensure comparability, we generated one random order of reviews once and used this order for generating $d_{labeled,\beta}$, such that $d_{labeled,\beta_1} \subset d_{labeled,\beta_2}$ for $\beta_1 \leq \beta_2$.

4.2 Model training

For applying neural network models for supervised fine-grained sentiment classification and also for MIL, it is necessary to determine a maximum number of tokens for each instance as well as a maximum number of instances for each review. We specified the maximum number of tokens for each instance by 60, resulting in only 1% of instances that had to be pruned to 60 tokens, as they had more tokens. Similarly, we specified the maximum number of instances for each review by 35, resulting in only 1% of reviews that had to be pruned to 35 instances, as they had more instances. For instances and reviews with shorter length, we used standard padding techniques to obtain a dataset, where all instances and reviews had the same lengths (i.e., 60 tokens per instance and 35 instances per review).

The hyperparameter of the used models were adopted from the works of Angelidis and Lapata (2018) and Devlin et al. (2019). More precisely, we used the pre-trained 'bert-base-cased' model from the transformer package[5] with an embedding size of 768 and trainable parameters. For the attention network, we used the BiGRU layer from the package PyTorch[6] with input size and hidden size equal to 768. We used the dropout rate 0.5 for the BiGRU layer and the log-softmax classification layer. The batch size for our evaluation was set to 4, with a learning rate of 0.00001 as proposed by Devlin et al. (2019). We trained all models for 10 epochs, and used the model with the epoch yielding the best performance on the validation set for assessing the performance on the test data. For training the models and for optimization of the loss functions, we used the standard optimizer AdamW from the package PyTorch.

4.3 Evaluation criteria

For evaluating the models for fine-grained sentiment classification on the test data, we use the F1 score, accuracy and root mean squared error (RMSE), which constitute common evaluation criteria for classification (e.g., cf. Sanyal et al., 2020). Since the focused task of fine-grained sentiment classification with classes $\{-1,0,1\}$ has more than two classes, we have to use the multi-class variants macro F1 score and weighted F1 score (cf. Blagec et al., 2021). While macro F1 score averages the F1 scores of the three classes, weighted F1 score computes a weighted average of the F1 scores of the three classes, weighted by the instances per class. Further, the accuracy measures the number of correctly predicted instances in relation to all instances. RMSE is also very appropriate, due to the ordered scale of sentiment classes, as it punishes high prediction errors (e.g., -1 instead of 1) more than small deviations (e.g., 0 and 1).

4.4 Results

To answer the second research question, we evaluate the MILPS model on different specifications of α and β. The results of this evaluation on the test data are given in Table 2. By construction of the loss function of the MILPS model, we remark that the supervised model is given by the MILPS model with $\alpha = 0$ and the MIL model is given by the MILPS model with $\alpha = 1$.

[5] https://huggingface.co/transformers/

[6] https://pytorch.org/

α	β	Weighted F1 Score	Macro F1 Score	Accuracy	RMSE
0	0%	0.19	0.18	0.36	0.80
1	0%	**0.51**	**0.52**	**0.57**	**0.79**
0	0.1%	0.51	0.52	0.57	0.79
0.25	0.1%	0.50	0.51	0.56	0.81
0.5	0.1%	**0.54**	**0.54**	**0.58**	**0.78**
0.75	0.1%	0.53	0.52	0.54	**0.78**
1	0.1%	0.32	0.27	0.42	1.07
0	1%	0.51	0.52	0.57	0.79
0.25	1%	**0.56***	**0.56***	**0.59***	**0.76***
0.5	1%	0.50	0.50	0.56	0.82
0.75	1%	0.51	0.50	0.53	0.85
1	1%	0.49	0.42	0.56	0.86

Table 2. Evaluation results of MILPS on different values of α and β. For a given β the best performance is indicated by bold font, while the asterisks highlight the best values over all values of α and β.

To better trace the effects of both parameters α and β and to anlyze the effect of a large dataset with fine-grained sentiment class information, the results with $\beta \in \{0.1\%, 1\%, 10\%\}$ are illustrated in Figure 1, depending on the values of α.

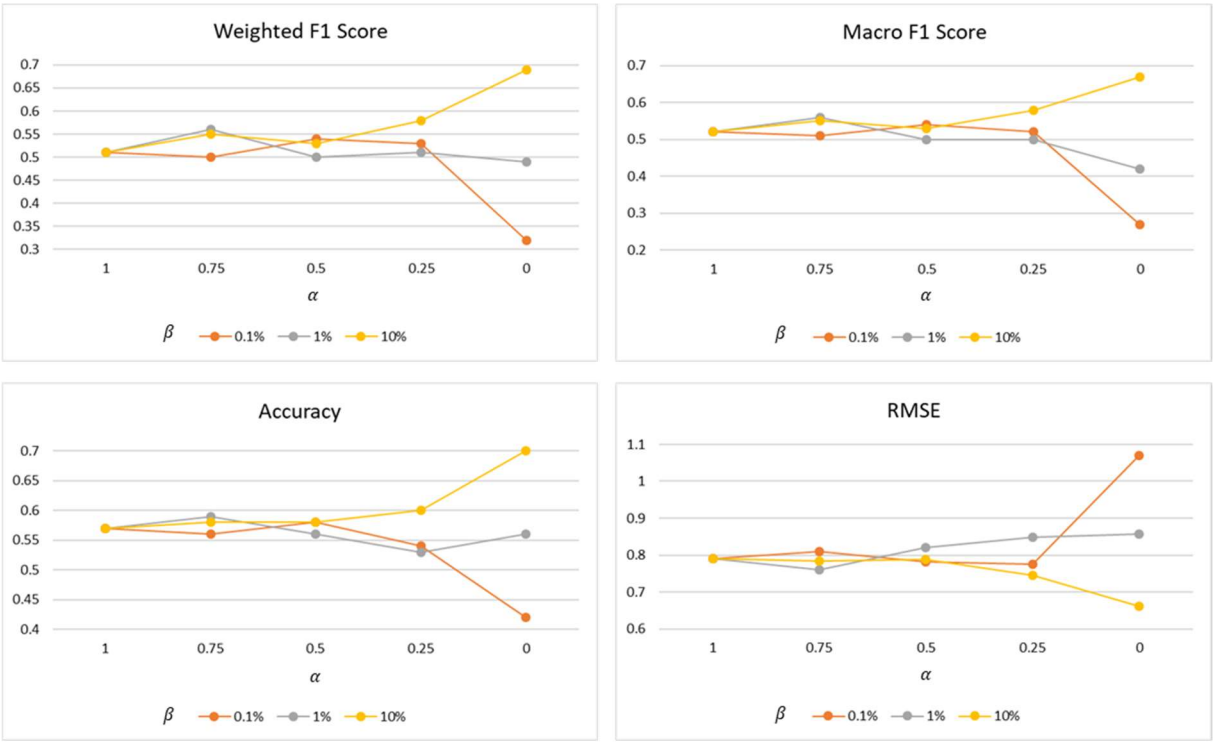

Figure 1. Effect of α (x-axis) and β (graphs) on performance of the model (y-axis).

Further, we also assess the performances for the individual sentiment classes -1,0 and 1 as outlined in Table 3. Here, we outline the F1 score for each specification of α and β for each sentiment class, as well as the averaged F1 score regarding α and β, which allows to analyze the effects of both parameters separately.

α	**0**	**0**	**0**	**0.5**	**0.5**	**0.5**	**1**	**1**	**1**
β	**0%**	**0.1%**	**1%**	**0%**	**0.1%**	**1%**	**0%**	**0.1%**	**1%**
negative	0	0	0	0.61	0.62	0.58	0.61	0.61	0.61
neutral	0.53	0.22	0.59	0.24	0.31	0.22	0.24	0.24	0.24
positive	0	0.58	0.67	0.71	0.7	0.7	0.71	0.71	0.71

α	**0**	**0.5**	**1**	averaged	averaged	averaged
β	averaged	averaged	averaged	**0%**	**0.1%**	**1%**
negative	0.00	0.60	0.61	0.41	0.41	0.40
neutral	0.45	0.26	0.24	0.34	0.26	0.35
positive	0.42	0.70	0.71	0.47	0.66	0.69

Table 3. Evaluation of effects of α and β on the F1 Score of each individual sentiment class.

5 Discussion and Managerial Implications

In this section, we discuss the results illustrated in Section 4.4 and give managerial implications.

First, our results in Table 2 confirm that MIL is already an appropriate and efficient way to predict fine-grained sentiment classes, which is indicated by the results with $\alpha = 1$ and $\beta = 0\%$. Compared to the trivial classifier ($\alpha = 0$ and $\beta = 0\%$), which always predicts the neutral class, the F1 scores and the accuracy of solely MIL are much higher. Only the values of RMSE are almost even, which is reasonable, as the neutral class predictions of the trivial classifier avoid higher errors (e.g., predicting 1 instead of the true class -1, resulting in a squared error of 4 compared to a squared error of 1 for predicting 0).

Second, the results in Table 2 show that for limited data with fine-grained sentiment class information (i.e., $\beta = 0\%$, 0.1% and 1% corresponding to 0, 7 and 74 sentences with sentiment labels) the proposed novel MILPS model yields the highest performance and outperforms the MIL model as well as the supervised classification. In particular, compared to the MIL, our extended MILPS model (with $\alpha = 0.25$ and $\beta = 1\%$), which also incorporates a few labeled sentiments, increases the F1 score by 5% on average. This is a very notable increase in performance in consideration of the circumstance that only 74 instances have to be labeled with fine-grained sentiments by human annotators, which takes approximately 2 hours of work. Compared to this small resource costs, the increase in performance by using MILPS is remarkable.

Third, the results illustrated in Figure 1 also contain the evaluation for $\beta = 10\%$ (corresponding to 737 labeled instances). This shows that utilizing solely the supervised model for fine-grained sentiment classification is best, when having a large amount of instances with fine-grained sentiment labels available. In this setting, the supervised model outperforms both, the MIL model as well as the proposed MILPS model. This is not surprising, as the supervised model is trained on exactly the same task as the model is evaluated on. In contrast, MIL and MILPS are also trained for predicting the review rating (i.e., review-level sentiment), which may reduce the focus on the aimed task of fine-grained sentiment classification. However, it is interesting that – even in this setting – the proposed MILPS model yields better results than the MIL model.

Fourth, due to hardware limitations regarding computational power in our evaluation, we could only use a dataset with 1,266 reviews. As the advantage of MIL and MILPS is by leveraging large volumes of review texts with associated star ratings to infer fine-grained sentiment classes, it can be assumed that the performance of both approaches can be increased significantly by utilizing a much larger volume of reviews, when businesses use these approaches, which have better hardware resources.

Fifth, we evaluated the performance of the models for each individual sentiment class based on $\alpha \in \{0,0.5,1\}$ and $\beta \in \{0\%, 0.1\%, 1\%\}$ (cf. Table 3). Here, it is interesting, that the supervised model ($\alpha = 0$) only predicts the classes neutral and positive, avoiding the classification of negatives (indicated by F1 scores equal to zero for the negative class). In contrast, the MIL model and the MILPS model generate predictions regarding all three sentiment classes, with high F1 scores for the positive and the negative class. This is especially important, as in sentiment analysis the focus is on instances with more extreme sentiments. Further, the F1 scores of the MILPS model (α=0.5, β averaged) is a little bit more balanced than the F1 scores of the MIL model (α=1, β averaged). In addition, the ratio of labeled instances β has almost no impact on the F1 score of the negative class, but with increasing β, the F1 score for the positive class increases. Overall, our results show that the proposed MILPS model, which extends MIL by means

of fine-grained supervision on (few) labeled instances, outperforms MIL model as well as supervised model for fine-grained sentiment classification in online consumer reviews, when only a few instances with fine-grained sentiment class information are available.

Our research has important managerial implications. By proposing MILPS model, we enable businesses as well as review platforms to conduct fine-grained sentiment analyses with good performance for online consumer reviews in a very cost-efficient way. In particular, fine-grained sentiment analysis of consumer reviews enables businesses to infer important consumer feedback and allows review platforms to model consumer preferences to improve personalization of their platforms (e.g., by personalized recommendations). For businesses and review platforms, it is essential to conduct fine-grained sentiment analysis, which yields good performance and is economical. This is especially important in vast changing market environments, as in such cases, it is very costly to generate large datasets with labeled fine-grained sentiment class information, if a change occurs. In particular, it would be necessary to generate costly new labeled datasets, when circumstances change (e.g., new products evolve on the market). Thus, our approach paves the way for an economical fine-grained sentiment analysis with high performance.

6 Conclusion

In this paper, we investigated the two research questions (i) how instance labels can be incorporated into MIL models to obtain MILPS models for fine-grained sentiment classification and (ii) whether this can improve the performance of fine-grained sentiment classification compared to MIL models and supervised models. Therefore, we proposed a novel MILPS model that extends MIL models with partly fine-grained supervision by incorporating instance sentiment labels (for only few instances), and compare the performance of a state-of-the-art supervised model, a state-of-the-art MIL model and the MILPS model on a dataset comprising online consumer reviews for restaurants. By our evaluation, we find that the MILPS models enable improved performance of fine-grained sentiment classification compared to the MIL model with highly reduced labeling efforts, which is therefore cost-efficient and economical.

Nevertheless, our work also has limitations, which could serve as starting points for future research. While we used a dataset comprising reviews for local businesses (e.g., restaurants, bars and cafés), it would be also interesting to assess the performance of MILPS on review datasets of other domains (e.g., electronic devices or movies), which could further validate the robustness and generalizability of our approach. Furthermore, our dataset was restricted to only 1,266 reviews due to hardware limitations. As pointed out in Section 5, the performance of MILPS and MIL unfolds its full potential by means of high-volume datasets. Therefore, it would be interesting to see works that apply the MILPS model to very large datasets, which would further substantiate the strengths of our proposed approach.

References

Angelidis, S. and M. Lapata (2018). “Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis” *Transactions of the Association for Computational Linguistics* 6, 17–31.

Blagec, K., G. Dorffner, M. Moradi and M. Samwald (2021). *A critical analysis of metrics used for measuring progress in artificial intelligence*.

Correia, J., I. Trancoso and B. Raj (2016). “Adaptation of SVM for MIL for inferring the polarity of movies and movie reviews”. In: *SLT 2016. 2016 IEEE Workshop on Spoken Language Technology : proceedings : December 13-16, 2016, San Diego, California, U.S.A*. Piscataway, NJ: IEEE, pp. 258–264.

Debortoli, S., O. Müller, I. Junglas and J. vom Brocke (2016). “Text mining for information systems researchers: An annotated topic modeling tutorial” *Communications of the Association for Information Systems* 39 (1), 7.

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North*. Ed. by J. Burstein, C. Doran, T. Solorio. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 4171–4186.

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

Hoang, M., O. A. Bihorac and J. Rouces (2019). “Aspect-based sentiment analysis using bert” *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187–196.

Jabr, W., B. Liu, D. Yin and H. Zhang (2020). *MIS Quarterly Research Curation on Online Word-of-Mouth Research Curation Team:* MIS Quaterly.

Kotzias, D., M. Denil, N. de Freitas and P. Smyth (2015). "From Group to Individual Labels Using Deep Features". In: ACM.

Levy, Y. and T. J. Ellis (2006). "A systems approach to conduct an effective literature review in support of information systems research" *Informing Science* 9.

Lutz, B., N. Pröllochs and D. Neumann (2019). "The Longer the Better? The Interplay Between Review Length and Line of Argumentation in Online Consumer Reviews". In: *Proceedings of the International Conference on Information Systems 2019*.

Nayak, G., R. Ghosh, X. Jia, V. Mithafi and V. Kumar (2020). "Semi-supervised Classification using Attention-based Regularization on Coarse-resolution Data". In C. Demeniconi and N. V. Chawla (eds.) *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 253–261. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Pannala, N. U., C. P. Nawarathna, J. T. K. Jayakody, L. Rupasinghe and K. Krishnadeva (2016). "Supervised Learning Based Approach to Aspect Based Sentiment Analysis". In: *2016 16th IEEE International Conference on Computer and Information Technology - CIT 2016, 2016 6th International Symposium on Cloud and Service Computing - IEEE SC2 2016, 2016 International Symposium on Security and Privacy in Social Networks and Big Data - SocialSec 2016*. *Proceedings: 7-10 December 2016, Nadi, Fiji*. Piscataway, NJ: IEEE, pp. 662–666.

Pappas, N. and A. Popescu-Belis (2014). "Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis". In: *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*, pp. 455–466.

Pappas, N. and A. Popescu-Belis (2017). "Explicit Document Modeling through Weighted Multiple-Instance Learning" *Journal of Artificial Intelligence Research* 58, 591–626.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. de Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra and G. Eryiğit (2016). "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Ed. by S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, T. Zesch. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 19–30.

Pontiki, M., D. Galanis, H. Papageorgiou, S. Manandhar and I. Androutsopoulos (2015). "Semeval-2015 task 12: Aspect based sentiment analysis". In: *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 486–495.

Sanyal, D., N. Bosch and L. Paquette (2020). "Feature Selection Metrics: Similarities, Differences, and Characteristics of the Selected Models" *International Educational Data Mining Society*.

Siering, M., A. V. Deokar and C. Janze (2018). "Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews" *Decision Support Systems* 107, 52–63.

van Engelen, J. E. and H. H. Hoos (2020). "A survey on semi-supervised learning" *Machine Learning* 109 (2), 373–440.

4 Explanations for RS and ABSA

"Certain things in life simply have to be experienced and never explained.
Love is such a thing."
Paulo Coelho (*1947)

4.1 Paper: Data Quality in Recommender Systems: The Impact of Completeness of Item Content Data on Prediction Accuracy of Recommender Systems

Current Status	**Citation**
This paper is **accepted** and **published** in Volume 31, Issue 2 in the journal *Electronic Markets*.	Heinrich, B., M. Hopf, D. Lohninger, A. Schiller and M. Szubartowicz (2021). "Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems" Electronic Markets 31 (2), 389–409.

Data Quality in Recommender Systems: The Impact of Completeness of Item Content Data on Prediction Accuracy of Recommender Systems

Bernd Heinrich, University of Regensburg, Regensburg, Germany, bernd.heinrich@ur.de
Marcus Hopf, University of Regensburg, Regensburg, Germany, marcus.hopf@ur.de
Daniel Lohninger, University of Regensburg, Regensburg, Germany, daniel.lohninger@ur.de
Alexander Schiller, University of Regensburg, Regensburg, Germany, alexander.schiller@ur.de
Michael Szubartowicz, University of Regensburg, Regensburg, Germany, michael.szubartowicz@ur.de

Abstract:

Recommender systems strive to guide users, especially in the field of e-commerce, to their individually best choice when a large number of alternatives is available. In general, literature suggests that the quality of data which a recommender system is based on may have important impact on recommendation quality. In this paper, we focus on the data quality dimension completeness of item content data (i.e., features of items and their feature values) and investigate its impact on the prediction accuracy of recommender systems. In particular, we examine the increase in completeness per item, per user and per feature as moderators for this impact. To this end, we present a theoretical model based on the literature and derive ten hypotheses. We test these hypotheses on two real-world data sets, one from two leading web portals for restaurant reviews and another one from a movie review portal. The results strongly support that, in general, the prediction accuracy is positively influenced by increased completeness. However, the results also reveal, contrary to existing literature, that among others increasing completeness by adding features which differ significantly from already existing features (i.e., a high diversity) does not positively influence the prediction accuracy of recommender systems.

Introduction

Recommender systems strive to guide users to their individually best choice when a large number of alternatives is available. Due to a broad variety of interesting problem settings for scholars and a plethora of practical applications, recommender systems continue to be a topic widely discussed in literature (Adomavicius and Tuzhilin 2005; Bobadilla et al. 2013; Karatzoglou and Hidasi 2017). For example, in recent years, many of these practical applications have been in the field of e-commerce and electronic markets (Li and Karahanna 2015; Lu et al. 2015; Ricci et al. 2011). Thereby, recommender systems “have become one of the most powerful and popular tools” (Ricci

et al. 2011), mainly because of the large amount of available data about items (e.g., songs or movies). Here, usually, a choice amongst an abundance of items needs to be made, which has inspired providers such as *Netflix* or *Spotify* to develop elaborate recommender systems (Bell et al. 2007; Gomez-Uribe and Hunt 2016; Song et al. 2013). Similarly, recommender systems can assist users in their choice of which restaurant to visit or in which hotel to stay (Levi et al. 2012; Vargas-Govea et al. 2011). In this context, several works suggest that the quality of the determined recommendations depends on the quality of the data which a recommender system is based on (Adomavicius and Zhang 2012; Felfernig et al. 2007; Konstan and Riedl 2012; Picault et al. 2011; Sar Shalom et al. 2015). As discussed by Jannach et al. (2016), these works mainly investigate the data quality of rating data (e.g., how to achieve the most accurate completion of the user-item matrix with rating predictions) and therefore, propose to leverage additional user data such as the user's context, the user's browsing history or the user's social graph. In contrast to these articles mainly discussing data quality of user or rating data (cf. Section "Background"), this paper focuses on data quality of *item content* data, which means, features of items such as *Genre* or *Actors* of movies and their feature values.

In general, data quality constitutes a multidimensional construct (Pipino et al. 2002; Wand and Wang 1996; Wang et al. 1995) comprising several dimensions such as correctness, completeness and currency of data (Batini and Scannapieco 2016; Heinrich et al. 2018b; Lee et al. 2002; Redman 1996). Some existing works investigate and assess the impact of data quality and its dimensions in decision making (Feldman et al. 2018; Heinrich and Hristova 2016). As recommender systems are an important category of decision support systems, especially in electronic markets, we aim to examine the impact of *item content* data and their quality on the determined recommendations. Here, capturing a more complete view of this item content data (i.e. more available features and feature values) is of particular relevance (Adomavicius and Tuzhilin 2005; Pazzani and Billsus 2007; Picault et al. 2011). After all, "some representations capture only certain aspects of the content, but there are many others that would influence a user's experience" (Lops et al. 2011). Hence, in this paper, we focus on the data quality dimension *completeness*.

Batini et al. (2009) summarize that completeness can be understood as the amount to which an available data view includes data describing the corresponding set of considered real-world objects (cf., e.g., also Ballou and Pazer 1985; Redman 1996). Following this definition, we aim to investigate the impact of completeness on recommendation quality, with completeness being the amount of available features and their feature values describing the set of items. For instance, the movie feature *Genre* has multiple feasible feature values such as *Comedy*, *Drama*, *Thriller* and so forth, while the restaurant feature *Cuisine* has multiple feasible feature values such as *Italian*, *American* or *Mexican*. Providers covering such domains typically assign such feature values to items in order to describe and emphasize their (special) characteristics and thus, allow a more complete view on these items.

Moreover, to assess the impact of completeness on recommendation quality, we examine the prediction accuracy, which is by far the most discussed quality measure in recommender systems literature (Shani and Gunawardana 2011). In this paper, prediction accuracy is assessed by the familiar evaluation measures Root Mean Squared Error (RMSE), Precision, Recall and F1-measure enabling a broad but also differentiated analysis of the results. To the best of our knowledge, no existing work analyzes the impact of the amount of available features or feature values (*completeness of item content data*) on prediction accuracy. Thus, we focus on the following two research questions:

RQa: *Does the amount of available item features influence the prediction accuracy of recommender systems?*

RQb: *Does the amount of filled up missing item feature values influence the prediction accuracy of recommender systems?*

We address these research questions by formulating ten hypotheses based on a theoretical model derived from the literature. Further, we test the statistical significance of these hypotheses by means of both a t-test and a moderated regression analysis concerning the impact of the amount of available item features and their feature values on prediction accuracy. The results show that completeness of the item content data generally has a significant positive impact on prediction accuracy. However, the results also reveal some findings which are contrary to statements in existing literature (Mitra et al. 2002; Tabakhi and Moradi 2015) stating that adding features with low diversity to a data set has less positive impact on prediction accuracy than adding features with high diversity.

Further, this research is also interesting for practitioners. For instance, the rapid development in e-commerce implies a swiftly increasing number of heavily competing web portals in electronic markets. Thus, increasing prediction accuracy by additional features and feature values may lead to competitive advantages for a portal. Furthermore, portals nowadays have their own individual data sets, which usually vary in their features and feature values for items, even for portals of the same domain (e.g., restaurants as items). Extending a data set with additional item content data from another data set (e.g., in case of a meta search portal) can be highly valuable for a recommender system as the two data sets may offer a differing and, when combined, more complete view of the items at hand. While portals offering a meta view exist (e.g., *trivago.com* compiles pricing data from various hotel portals), these portals usually simply juxtapose the data and do not use it to provide recommendations based on additional features and feature values. Analyzing the impact of increased completeness of item content data on prediction accuracy may reveal substantial unused potential in this context.

The remainder of the paper is organized as follows: In the next section, we discuss related work regarding data quality in the context of recommender systems, especially in terms of the dimension completeness, and outline the theoretical model which is used to substantiate the hypotheses presented in the following section. Thereafter, we

discuss the used evaluation measures and testing methodology. In the evaluation section, we statistically test the significance of our hypotheses based on two different real-world data sets. Afterwards, we analyze and discuss the results and give some further practical implications. Finally, we summarize our work and point out limitations as well as directions for future research.

Background and Theoretical Model

This section consists of two subsections covering the literature background and the theoretical model for our research.

Background

In this subsection, we firstly analyze existing works related to our research questions. Thereafter, we identify the research gap which is addressed in this paper. Following the guidelines of standard approaches to prepare the related work (e.g., Levy and Ellis 2006), we performed a literature search on the databases ACM Digital Library, AIS Electronic Library, IEEE Xplore, ScienceDirect and Springer as well as the proceedings of the European and International Conference on Information Systems, the ACM Conference on Recommender Systems and the International Conference on Information Quality. The resulting papers were examined based on title, abstract and keywords, leading to thirteen remaining papers. We performed an additional forward and backward search on these papers, leading to a total of twenty-seven relevant papers. These papers were analyzed in detail and could be organized within three categories A, B and C. Works of category A discuss data quality issues in the context of recommender systems, whereas works of category B present recommender systems which deal with a data set extended by using web data sources. Works of category C investigate the impact of data characteristics such as the entropy of the distribution of rating data on recommendation quality. In the following, we discuss the relevant papers of each category.

The eight works in category A explicitly recognize the importance of data quality for recommender systems from a *general* perspective (Amatriain et al. 2009; Berkovsky et al. 2012; Burke and Ramezani 2011; Konstan and Riedl 2012; Lathia et al. 2009; Levi et al. 2012; Pessemier et al. 2010; Sar Shalom et al. 2015), including several approaches that deal with data quality issues. For instance, as data sparsity and inaccuracy have been identified to influence recommendation quality, Lathia et al. (2009) suggest to choose data sources for the application of a recommender system user-dependently. Sar Shalom et al. (2015) tackle sparsity and redundancy issues by deleting or omitting certain users or items while Pessemier et al. (2010) analyze consumption data such as ratings in regard to currency. Further, Levi et al. (2012) use text mining on user reviews from various sources to alleviate the cold start

problem of new users by assigning them to so-called context groups.

The four works in category B (implicitly) investigate completeness in recommender systems (Abel et al. 2013; Bostandjiev et al. 2012; Kayaalp et al. 2009; Ozsoy et al. 2015). More precisely, these works propose to use data from additional web sources to gain an extended data set and to increase recommendation quality in this way. Abel et al. (2013) study user profiles based on aggregated data sets from the social web and show that recommendation quality is improved by user profiles extended through several cross-system user-modelling strategies. Ozsoy et al. (2015) argue that recommendations can be improved by consolidating user data from multiple sources. In their experiments, they show that using multiple user features from several social networks produces an enhanced perspective of user behavior and preferences, leading to improved recommendations. Kayaalp et al. (2009) present an event recommender system for users of a social network. This system collects heterogeneous event data from various web pages to achieve an extended data set and proposes event recommendations on this basis. A further approach is proposed by Bostandjiev et al. (2012). They suggest to use multiple data sources such as Twitter, Facebook and Wikipedia to apply an individual recommender system on each data source. Afterwards, the recommendation results are combined aiming to improve recommendation quality.

The fifteen works in category C examine the impact of data characteristics (so-called meta-features) on recommendation quality. In particular, these works investigate the impact of data characteristics of rating data (e.g., Adomavicius and Zhang 2016; Griffith et al. 2012; Matuszyk and Spiliopoulou 2014), content data (Fortes et al. 2017) and other data such as binary purchase data (Geuens et al. 2018), social network graph data (Olteanu et al. 2014) or folksonomy data (Doerfel et al. 2016) on different performance measures of recommender systems. For instance, Cunha et al. (2016), Ekstrand and Riedl (2012) and Huang and Zeng (2005) aim to select the best recommender algorithm depending on data characteristics such as the entropy of ratings. Furthermore, Adomavicius and Zhang (2012), Basaran et al. (2017) and Grčar et al. (2006) analyze the recommendation quality based on rating data specific meta-features such as the user-item ratio. As meta-features usually provide valuable information, for instance, Sergis and Sampson (2016) and Zapata et al. (2015) enhance hybrid recommender systems by including the meta-features directly as input to the recommender algorithm.

Given this discussion, none of the works above investigates the impact of completeness of item content data on recommendation quality. The works in category A focus on data quality issues in recommender systems, analyzing the impact of dimensions such as accuracy and currency on recommendation quality. We extend this category of works by contributing investigations for the impact of completeness on recommendation quality. The works in category B focus on completeness aspects in the context of recommender systems. Abel et al. (2013) and Ozsoy et al. (2015) aim to improve recommendation quality by using more complete *user* data. Kayaalp et al. (2009) and

Bostandjiev et al. (2012) use multiple sources for data concerning items in the context of recommender systems. Here, Kayaalp et al. (2009) focus on the technical challenges arising from the integration of heterogeneous event data types for recommender systems and do not discuss the impact of completeness of item content data on recommendation quality. Bostandjiev et al. (2012) apply different recommender systems on each data source separately. Their resulting recommendation is the aggregation of the recommendations based on each single data source. Therefore, works in category B do not aim at an explanatory analysis or refer to a theoretical model to study whether recommendation quality is influenced by adding features and feature values. The works in category C focus on the impact of data characteristics on recommendation quality. While the majority of works study impact of data characteristics (meta-features) of rating data, only Fortes et al. (2017) investigate data characteristics in relation to item content data. They enhance the recommender system by including these data characteristics directly in the recommender algorithm as they aim for a predictive analysis. In contrast to the discussed works, which either focus on the consideration of *rating* data characteristics (e.g., entropy of rating distribution) or generate recommendations in a predictive analysis, we extend this category of works in two ways. Firstly, we explicitly investigate the impact of completeness of *item content* data on prediction accuracy. Secondly, we conduct an explanatory analysis based on causal hypotheses and a theoretical model, which strongly differs from predictive analytics (Shmueli and Koppius 2011). Both aspects have important implications in practice as the actual relevance of increasing the amount of available features and feature values for prediction accuracy is examined.

Theoretical Model

This subsection presents a theoretical model constituting a basis for the hypotheses discussed in the subsequent section. Research in the field of data quality shows an increasing tendency to study the impact of data quality of data views and data values (independent variable) on different evaluation criteria of decision support systems such as decision quality or data mining outcome (dependent variable) (e.g., Bharati and Chaudhury 2004; Blake and Mangiameli 2011; Feldman et al. 2018; Ge 2009; Woodall et al. 2015). More precisely, Blake and Mangiameli (2011) analyze the impact of the data quality dimensions accuracy, completeness, consistency and currency on data mining results in order to support decision-making. Woodall et al. (2015) investigate the impact of completeness on classification outcomes used for supporting users in their decision process. Bharati and Chaudhury (2004) examine the effects of accuracy, completeness and currency on the ability of an online analytical processing system to sustain decision-making. Ge (2009) focuses on accuracy, completeness and consistency and their impact on decision quality. Feldman et al. (2018) propose an analytical framework to investigate the impact of incomplete data sets on a binary classifier that serves for decision support.

The focus of these papers is to investigate in which way and to what extent the quality of data views and data values, especially the dimension completeness, influences evaluation criteria such as data mining outcome of particular decision support systems. Because recommender systems are a relevant category of decision support systems, especially in electronic markets, assisting users that face decision-making problems (Porcel and Herrera-Viedma 2010; Power et al. 2015), we derive the theoretical model from these works to examine the impact of completeness of item content data on prediction accuracy of recommender systems. Figure 1 presents this theoretical model.

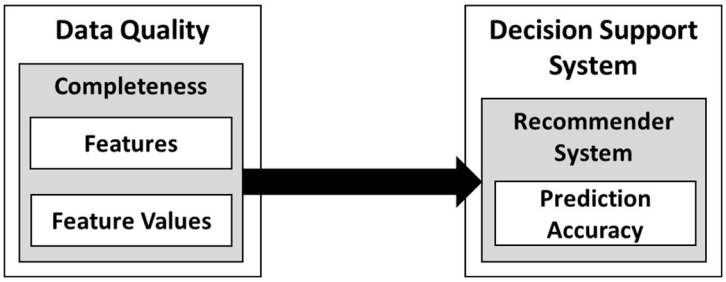

Figure 1. Theoretical Model

In the context of decision support systems, completeness is a frequently investigated dimension of data quality (Blake and Mangiameli 2011; Feldman et al. 2018; Ge 2009; Woodall et al. 2015). These works refer to completeness as the amount of available data views and data values. We take up this idea in the theoretical model and consider completeness by the amount of features and their feature values (cf. left side of Figure 1). As discussed above, features such as *Cuisine* can have multiple feasible feature values such as *Italian*, *American* or *Mexican*, which are assigned to items in order to describe and underline their characteristics enabling a more complete view on these items. Therefore, we focus on such features and their feature values when analyzing completeness. Similar to Bharati and Chaudhury (2004) and Ge (2009), the presented theoretical model in Figure 1 indicates a direct relation between data quality and evaluation criteria of decision support systems. In particular, the theoretical model suggests this relation between completeness of item content data and prediction accuracy of recommender systems (cf. right side of Figure 1). This model constitutes the foundation for the following hypotheses and is customized by different moderator variables to allow for a detailed analysis.

Hypotheses

Based on the theoretical model, we present ten hypotheses to address our research questions. Each hypothesis examines the impact of completeness of item content data on prediction accuracy from a different angle. Figure 2 at the end of this section shows an overview of all hypotheses.

Content-based and hybrid recommender systems, two major categories of recommender systems (Ning et al. 2015),

operate on item content data to propose items to users that they are likely to be interested in (Lops et al. 2011). For this kind of data, increased completeness means that more features and/or more feature values are assigned to items (cf. Section "Theoretical Model"). Thus, increased completeness in this sense can be achieved in two ways: First, by adding features and their feature values to the feature set. For instance, a feature *Actors* can be added to the feature set for the movie domain. Second, by filling up missing feature values. For example, an already available feature *Parking Information* stating the parking options of a restaurant may have missing values for some restaurants which can be filled up. This can be done in various ways, for example by surveys, analyses or imputation (cf. Section "Description and Preparation of Data Sets"). Hence, all following hypotheses address both ways of increasing completeness in correspondence with our research questions *RQa* and *RQb*. Hypotheses labelled "a" focus on completeness increased by adding features and their feature values, whereas hypotheses labelled "b" focus on completeness increased only by filling up missing feature values. For both types of hypotheses, we test whether an increase in prediction accuracy can be observed.

This discussion leads to the following first two hypotheses:

H1a: Adding features and their feature values leads to higher prediction accuracy.

H1b: Filling up missing feature values leads to higher prediction accuracy.

Hypothesis H1a pursues the idea that the preferences of users can be analyzed in more detail when more item features and their feature values are available and suggests that the prediction accuracy (assessed by RMSE, Precision, Recall and F1-measure; cf. Section "Assessing Prediction Accuracy") is thus higher. Hypothesis H1b follows the expectation that recommendations are more accurate when missing values of item features are filled up.

Depending on the analysis of Hypotheses H1a/b, it is further interesting whether the extent of increased completeness measured *per item*, *user* or *feature* influences the extent of increased prediction accuracy. Regarding items and users, this can be described more precisely as follows: Does the increase in the amount of additional features and feature values (type "a") or the increase in the amount of filled up feature values (type "b") positively moderate the impact of completeness on prediction accuracy for an item or a user?

Therefore, it is meaningful to examine moderator effects regarding users and items on the relationship between completeness and prediction accuracy. This discussion leads to further hypotheses, which consider the increase in the amount of additional features and feature values, respectively, the increase in the amount of filled up feature values, per item or per user. Beginning with items, we examine the following hypotheses:

H2a: The increase in the amount of additional features and their feature values *for an item* constitutes a

positive moderator on the impact of completeness on prediction accuracy.

H2b: The increase in the amount of filled up feature values *for an item* constitutes a positive moderator on the impact of completeness on prediction accuracy.

Analogously, we formulate the hypotheses regarding the increase in completeness for users as follows:

H3a: The increase in the amount of additional features and their feature values *regarding a user* constitutes a positive moderator on the impact of completeness on prediction accuracy.

H3b: The increase in the amount of filled up feature values *regarding a user* constitutes a positive moderator on the impact of completeness on prediction accuracy.

Similar to items and users, it appears reasonable that the extent of increased completeness *per feature* also influences the extent of increase in prediction accuracy. Consequently, the following hypotheses examine the moderator effect regarding features on the relationship between completeness and prediction accuracy.

At first, we focus on a higher *amount* of values of added or filled up features, respectively, which leads to the following two hypotheses:

H4a: The increase in the amount of feature values *for an additional feature* constitutes a positive moderator on the impact of completeness on prediction accuracy.

H4b: The increase in the amount of feature values *for a filled up feature* constitutes a positive moderator on the impact of completeness on prediction accuracy.

Finally, we focus on increased completeness through higher *diversity* of added or filled up features. Additional features may have similar feature value assignments for items as already existing features. In particular, adding a feature, which has exactly the same feature values for items as an existing feature, may not influence the prediction accuracy at all, since such a feature does not add any further diversity to the item content data (Mitra et al. 2002; Tabakhi and Moradi 2015). In contrast, adding features that provide a high diversity to the item content data enhance the recommender system's ability to differentiate items and users and thus may lead to a high increase in prediction accuracy. Therefore, we consider the following hypotheses expecting a moderator effect when adding or filling up features depending on their diversity:

H5a: The diversity *for an additional feature* constitutes a positive moderator on the impact of completeness on prediction accuracy.

H5b: The diversity *for a filled up feature* constitutes a positive moderator on the impact of completeness

on prediction accuracy.

Figure 2 customizes the theoretical model (cf. Figure 1) by incorporating moderator variables and the stated hypotheses. In general, it shows the expected impact of the data quality dimension completeness on the prediction accuracy as stated by Hypotheses H1a/b. Additionally, it illustrates the hypotheses examining a moderating effect for the increase in completeness per item (Hypotheses H2a/b), per user (Hypotheses H3a/b) and per feature (Hypotheses H4a/b, H5a/b).

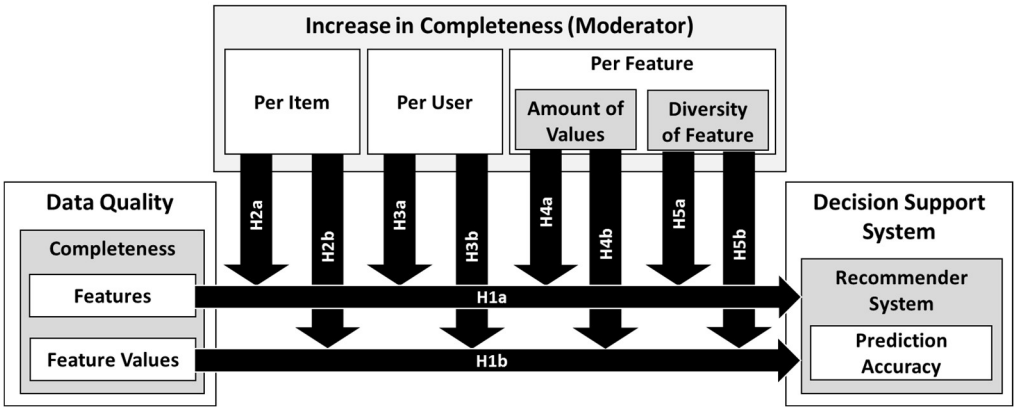

Figure 2. Overview for Hypotheses H1-H5

Methodology

In this section, we introduce the models used to test Hypotheses H1-H5. To do so and to assess prediction accuracy as the dependent variable, we first discuss selected measures which allow differentiated analyses and interpretations regarding the impact on prediction accuracy. Thereafter, we describe the testing methodology for Hypotheses H1a/b as well as the regression models for testing Hypotheses H2-H5.

Assessing Prediction Accuracy

To enable a detailed and careful analysis of the results of the Hypotheses H1-H5, we assessed prediction accuracy by means of different measures from literature, namely RMSE, Precision, Recall and F1-measure (Gunawardana and Shani 2015). RMSE as shown in Equation (1) is one of the most popular measures for assessing prediction accuracy (Gunawardana and Shani 2015) and is defined by the term

$$RMSE = \sqrt{\frac{1}{|T|} \cdot \sum_{(u,i) \in T} (\hat{r}_{ui} - r_{ui})^2}, \quad (1)$$

where T is a test set of user-item pairs (u, i) for which the ratings $\hat{r}_{ui}$ are predicted by the recommender system and the actual ratings r_{ui} are known. RMSE received special attention by the Netflix Prize Challenge in 2006 (Koren

2009). Its main characteristic is that higher errors (i.e., the difference between predicted and actual rating) are weighted stronger through its quadratic structure than lower errors. Further, usually the predicted ratings $\hat{r}_{ui}$ are continuous (real-valued) and the actual ratings r_{ui} are discrete (and ordered). Hence, minor RMSE value changes may not result in a different mapping (by rounding) of the continuous predicted rating $\hat{r}_{ui}$ to a discrete star rating $\widehat{dr}_{ui} \in \{1, \dots, 5\}$. This means that the mapping to a discrete star rating may not change, even with an improved RMSE value. Therefore, it is also necessary to assess whether the mapping of continuous predicted ratings $\hat{r}_{ui}$ to discrete star ratings $\widehat{dr}_{ui}$ changes or improves with the increase in completeness and the expected increase in prediction accuracy. To evaluate this, Precision, Recall, and F1-measure are the most important measures. These measures assess whether or not the predicted rating level $\widehat{dr}_{ui}$ exactly coincides with the actual true rating level r_{ui} for each user-item pair (u, i) (Aggarwal 2014). Precision and Recall are calculated as the average of the Precision and Recall values for each star rating level $k \in \{1,2,3,4,5\}$, which are given by the following terms.

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{2}$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{3}$$

Here, TP_k is the number of user-item pairs (u, i) with $r_{ui} = k$ and $\widehat{dr}_{ui} = k$ ("true positives"), FP_k as shown in Equation (2) is the number of user-item pairs (u, i) with $r_{ui} \neq k$ and $\widehat{dr}_{ui} = k$ ("false positives"), and FN_k as shown in Equation (3) is the number of user-item pairs (u, i) with $r_{ui} = k$ and $\widehat{dr}_{ui} \neq k$ ("false negatives"). F1-measure as shown in Equation (4) is then given by the harmonic mean of Precision and Recall

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{4}$$

The main difference in interpretation of these measures is that the Precision, Recall and F1-measure focus on correct or incorrect mappings of predicted and actual star ratings while ignoring the (real-valued) error size, which is in the focus of RMSE.

Model for Hypotheses H1a/b

Each of the Hypotheses H1a and H1b focuses on a comparison of the prediction accuracy of two item content data sets, one data set without increased completeness and the other data set with increased completeness (cf. Figure 3). In both cases, we initially do not consider any moderator variable. To test the significance of both hypotheses, we used the paired Student's t-test, a broadly applied test in the evaluation of recommender systems to compare the results of two different settings, while in both settings the considered set of user ratings remains the same (Shani and

Gunawardana 2011). More precisely, the t-test was used to compare each of the measures RMSE, Precision, Recall and the F1-measure (and thus the prediction accuracy) based on the data set with increased completeness (i.e., when adding features and their feature values or when filling up missing feature values) and based on the data set without increased completeness.

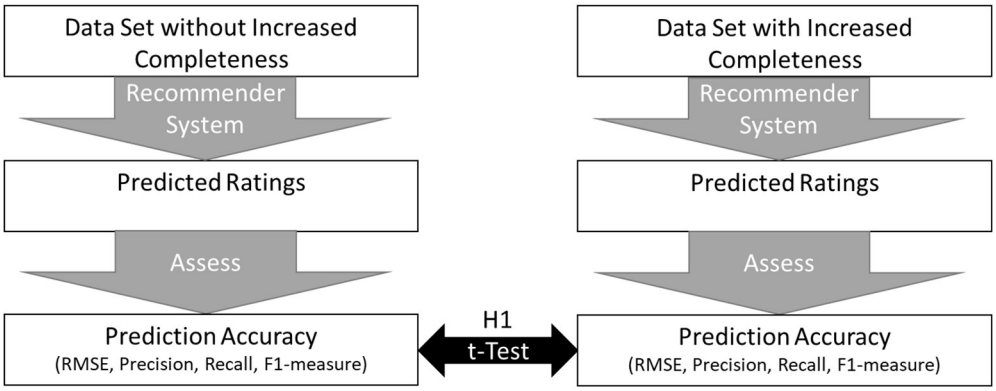

Figure 3. Testing Hypotheses H1a/b

Model for Hypotheses H2-H5

The Hypotheses H2-H5 analyze whether the increase in completeness per item, user or feature moderates the impact of completeness on the increase in prediction accuracy caused by adding features and their feature values (hypotheses of type "a") or by filling up missing feature values (hypotheses of type "b"). This means that the tests of the Hypotheses H2-H5 are organized in a similar way. Therefore, we describe the general structure for all of these tests in the following.

To test moderator effects on the impact of completeness on increased prediction accuracy, we chose moderated regression analysis (cf., e.g., Cohen et al. 2003; Dawson 2014; Hayes 2013; Helm and Mark 2012) as it is a widespread statistical tool to test whether the relationship between two variables is dependent on a third variable (the moderator). The underlying regression model is represented by the equation

$$y = b_0 + b_1 \cdot x + b_2 \cdot z + b_3 \cdot x \cdot z. \quad (5)$$

Here, y is the dependent or endogenous variable (criterion), x is the independent or exogenous variable (predictor) and z is the moderator variable. Regarding Hypotheses H2-H5, the endogenous variable y constitutes the (expected) increase in prediction accuracy measured by RMSE, Precision, Recall and F1-measure while the exogenous variable x indicates whether the data set with increased completeness or the data set without increased completeness is used. The moderator variable z constitutes the increase in completeness. More precisely, for H2a, H3a and H4a, the variable z represents the increase in additional features and feature values and for H2b, H3b and H4b, the variable z

represents the increase in filled up feature values. Similar, for H5a, the variable z represents the diversity of added features, and for H5b, the variable z represents the diversity of filled up features.

Besides the common interpretation of the coefficient b_0 as well as the coefficients b_1 and b_2 (*first order effects* of the regression model), the product term $x \cdot z$ and its coefficient b_3 are of special interest. This term represents the interaction (moderation) of two variables. More precisely, the coefficient b_3 estimates how much the slope of x changes as z changes. This represents how much the impact of increased completeness on prediction accuracy is influenced by the (different) values of the moderator variable. Therefore, a hypothesis proposing a moderator effect can be supported, if there is evidence that b_3 is different from zero with a certain level of significance.

In case of a moderator effect, the strength of this effect can be assessed by Cohen's f^2. Here, the coefficient of determination regarding the regression model depicted in Equation (5) is compared to the coefficient of determination of the regression model without the interaction term, which means,

$$y = b_0 + b_1 \cdot x + b_2 \cdot z. \tag{6}$$

Denoting the coefficient of determination R^2 according to each Equation (5) and (6) (i.e., R_1^2 and R_2^2), Cohen's f^2 is given by the term

$$f^2 = \frac{R_1^2 - R_2^2}{1 - R_1^2}. \tag{7}$$

Cohen's f^2 measures the relative increase in the explained variance of y when adding the interaction term to Equation (6) as shown in Equation (5). In Cohen (1988) the values 0.02, 0.15 and 0.35 are suggested for f^2 to indicate small, medium or large moderator effect sizes, which is critically discussed in scientific literature (Aguinis et al. 2005; Gignac and Szodorai 2016; Helm and Mark 2012). For instance, Aguinis et al. (2005) conducted a review of 261 articles published in several journals (maintaining high methodological standards) in order to analyze the size of moderating effects. They found that the mean of Cohen's f^2 was about 0.009 (with a standard deviation of 0.025), and the median about 0.002 with a positively skewed distribution (skewness = 6.52). This indicates that – regarding the suggested values of Cohen (1988) – a medium or strong moderator effect can be rarely attained. In their discussion, they encourage researchers to "plan future research designs based on smaller (and more realistic) targeted effect sizes" (Aguinis et al. 2005) as long as the observed effect has a meaningful impact and interpretation for science and practice.

Evaluation

In this section, we outline the test procedure and results of our empirical evaluation. Initially, we describe both used real-world data sets. Afterwards, we introduce the recommender system which was applied to these data sets and outline in detail how we tested each hypothesis. We conclude the section by presenting the results of these tests.

Description and Preparation of Data Sets

For testing our hypotheses, we prepared two real-world data sets. While the first data set contains a large number of user-generated ratings about restaurants and was retrieved from two leading advertising web portals, the second data set is based on the non-commercial movielens data set containing approximately one million ratings (Harper and Konstan 2015). In both data sets, the ratings are assessments of items by users and hence, each rating corresponds to exactly one user and one item. Further, the rating values are given on an ordinal, five-tier scale of stars, ranging from 1 star to 5 stars.

Restaurant Data Set

In the first data set, one portal (Portal 1) focuses on local businesses such as bars or restaurants and provided over 100 million ratings by 2018. The second portal (Portal 2) specializes on travel opportunities and businesses such as restaurants providing over 400 million ratings by 2018. Since each web portal provided a vast amount of data, we focused on an excerpt and chose rating data of restaurants from the area of New York City, USA, because the high number of restaurants in this area allows for testing each hypothesis on a sufficiently high number of items, users or features, respectively. This led to a data set with more than 2.2 million ratings provided by over 550,000 users on more than 18,500 restaurants from Portal 1 and more than 720,000 ratings from about 375,000 users for more than 8,600 restaurants from Portal 2. Table 1 describes the restaurant data set.

	Portal 1	**Portal 2**
# of Users	556,462	374,960
# of Restaurants	18,507	8,631
# of Ratings	2,252,224	721,416

Table 1. Description of the Restaurant Data Set

Both web portals provide features such as *Cuisine* with multiple feasible feature values such as *Italian*, *American* or *Mexican*. In both portals, these feature values are assigned to an item. Other features of restaurants are *Special Diets* with feature values such as *Vegetarian, Vegan* or *Gluten-free* and *Type of Establishment* with feature values such as *Café, Bistro* or *Bar*. With this in mind, the knowledge about feature value assignments is especially relevant for each

item in this data set. In the case that a feature value is unknown, we indicated the missing feature value by the value *N/A* (not available).

From Portal 1 we retrieved an item content data set with 13 different features, denoted by P1, while Portal 2 provided an item content data set with 12 different features, denoted by P2. As only Portal 1 yielded features containing missing values, we split up P1 into an item content data set P1.1, containing only the seven features without missing values, and an item content data set P1.2, containing only the six features with missing values. More precisely, 44% of all possible 425,661 feature values for the six features of P1.2 were not available for the 18,507 restaurants of Portal 1. Table 2 illustrates the features and feature values per portal.

	Portal 1		**Portal 2**
Item Content Data Set	**P1**		**P2**
	P1.1	**P1.2**	
# of Features	7	6	12
# of Missing Feature Values	0 (0%)	189,164 (44%)	0 (0%)

Table 2. Features and Feature Values provided by the two Web Portals of the Restaurant Data Set

Data sets for hypotheses of type "a"

To prepare the data set for testing the hypotheses of type "a", we focused on the features from P1.1 from Portal 1 and P2 from Portal 2 that did not contain any missing data. This was important in order to carefully separate hypotheses of type "a" and of type "b". To obtain the joint feature set for a restaurant from the item content data sets P1.1 and P2, it was necessary to match restaurants between both portals. We thus conducted record linkage, which is the task of identifying records that refer to the same entity across different data sources (Christen 2012). To do so, we used a common rule-based classification model. The model was built using manually labelled training data and evaluated by quality measures. The classification resulted in 5,367 restaurants matching across the two portals with a false discovery rate below 1% on manually labelled test data. This means that less than 1% of these restaurants were incorrectly classified as matching. We exclusively focused on such matching restaurants to test the hypotheses of type "a" because these restaurants had added features compared to the features in each single portal. Furthermore, for each portal, we considered users with more than 30 ratings in order to only evaluate users with a substantial number of ratings (Sarwar et al. 2002). To increase completeness, features from Portal 2 were added to the feature set of Portal 1 and vice versa. This resulted in two cases used for testing the hypotheses of type "a": The data for the first case originated from Portal 1, consisted of 5,367 items with 367,182 ratings of 8,138 users and was evaluated using the item content data sets P1.1 as baseline and P2 as set of additional features and their feature values. The data for the second case originated from Portal 2, comprised the same 5,367 items with 20,659 ratings of 505 users and was evaluated using the item content data sets P2 as baseline and P1.1 as set of additional features (cf. Table 3).

Data sets for hypotheses of type "b"

To prepare data for testing the hypotheses of type "b", we focused on the first portal, as the second portal did not provide any features with missing values. In this case, to fill up missing feature values in the item content data set P1.2 containing six features, we used the common nearest neighbor imputation technique (Enders 2010). Similar to above, this imputation was evaluated by means of training and test data as well as quality measures. Missing values were imputed with a mean absolute error of only 0.299 for the test data. Again, we considered users with more than 30 ratings. This led to the data for testing the hypotheses of type "b" consisting of 18,507 restaurants with 731,395 ratings of 10,556 users, which was evaluated comparing the item content data sets P1.2 as baseline (consisting of 236,497 feature values) and P1.2' as set of baseline features with filled up feature values (consisting of 425,661 feature values including the 189,164 filled up feature values) (cf. Table 3).

	Hypotheses of Type "a" originating from Portal 1		**Hypotheses of Type "a"** originating from Portal 2		**Hypotheses of Type "b"** originating from Portal 1	
Item Content Data Set	**P1.1** (Baseline)	**P1.1&P2** (Baseline & add. features)	**P2** (Baseline)	**P1.1&P2** (Baseline & add. features)	**P1.2** (Baseline)	**P1.2'** (Baseline & filled up feature values)
# of Features/ # of Feature Values	7	19	12	19	236,497	425,661
# of Items	5,367		5,367		18,507	
# of Ratings	367,182		20,659		731,395	
# of Users	8,138		505		10,556	

Table 3. Description of the Data Bases for Evaluating Hypotheses H1a/b-H5a/b on the Restaurant Data Set

Movie Data Set

The second data set focuses on movies and originates from the research lab grouplens, which provides data sets with up to 20 million ratings from the non-commercial web portal movielens by 2016. Since the movielens data sets have been updated since 1998, new features and feature values have been added in new versions. To enable an evaluation based on a larger amount of ratings, we consider the data set from 2003 with only one feature and its most recent version from 2016 with five additional features and their feature values. The old version (OldV) of the movielens data set from 2003 contains over one million ratings provided by over 6,000 users on approximately 3,900 movies, while the new version (NewV) consists of over 20 million ratings from about 140,000 users for more than 27,000

	OldV	**NewV**
# of Users	6,040	138,493
# of Movies	3,883	27,278
# of Ratings	1,000,209	20,000,263

Table 4. Description of the Movie Data Set

movies. Table 4 describes the movie data set. Similar to the restaurant data set, both versions of the movielens data set provide the feature *Genre* with multiple feasible feature values such as *Comedy*, *Drama* or *Thriller*, while the new version provides additional features and their feature values such as *Actors* and *Country of Origin* each with according feature values. For example, the additional feature *Actors* in the version NewV indicates the top billed actors of the movie cast. Both versions do not yield features containing missing values, which means that only hypotheses of type "a" could be tested on the movie data set. Table 5 illustrates the features and feature values per version.

Item Content Data Set	**OldV**	**NewV**
# of Features	1	6
# of Missing Feature Values	0 (0%)	0 (0%)

Table 5. Features and Feature Values provided by the two Versions of the Movie Data Set

Data set for hypotheses of type "a"

Since the movie data set consists of an old and a new version, it is clear that the baseline item content data set is given by the old version and the item content data set with increased completeness is given by the union of both versions. Similar to the restaurant data set, the joint feature set for a movie was obtained by matching movies between both versions. As the movielens identifiers of the movies did not change between both versions (except from 24 movies, which were removed), record linkage was easy to conduct. Furthermore, the 6,040 users in both versions had at least 20 ratings, enabling a substantial number of ratings for the evaluation. Since 24 movies and their corresponding 2,175 ratings had been removed in the new version NewV, this resulted in content data sets consisting of 3,859 items with 998,034 ratings of 6,040 users and was evaluated using the item content data sets OldV as baseline and NewV as set of additional features and their feature values (cf. Table 6).

	Hypotheses of Type "a" originating from the old version of the movielens data set	
Item Content Data Set	**OldV** (Baseline)	**OldV&NewV** (Baseline & add. features)
# of Features	1	6
# of Items	3,859	
# of Ratings	998,034	
# of Users	6,040	

Table 6. Description of the Data Bases for Evaluating Hypotheses H1a-H5a on the Movie Data Set

Used Recommender System

For our evaluation, we used the hybrid recommender system approach *Content-Boosted Matrix Factorization* (CBMF) as presented by Forbes and Zhu (2011) and Nguyen and Zhu (2013). Matrix factorization approaches

became very popular by the contest on the Netflix Grand Prize, which started 2006 and ended 2009 (Koren et al. 2009; Koren 2009). They are now state-of-the-art models in the research of recommender systems (Kim et al. 2016; Ning et al. 2017; Symeonidis 2016). CBMF is able to utilize both non-content data (ratings) and, in particular, content data (features and feature values of items). Like all matrix factorization models, CBMF models are learned by optimization and therefore, preliminary steps such as feature weighting or feature selection are not necessary for CBMF (Koren et al. 2009; Nguyen and Zhu 2013).

CBMF learns a d-dimensional vector of latent factors $p_u \in \mathbb{R}^d$ for each user u and a d-dimensional vector of latent factors $a_f \in \mathbb{R}^d$ for each feature f, such that the actual rating r_{ui} for a user-item pair (u, i) is approximated by the predicted star rating $\hat{r}_{ui} = p_u^T q_i$, with $q_i = \sum_{f \in F_i} a_f$ and F_i being the set of features that are assigned to item i. In our evaluation, we used the default configuration for CBMF as described, for instance, by Nguyen and Zhu (2013). Excepting this default configuration concerns the regularization penalty factor λ, which has to be adjusted depending on the data set (Koren et al. 2009). Thus, to determine this factor we conducted cross-validation tests as described by Koren et al. (2009). For instance, the value $\lambda = 10 \times 10^{-6}$ (cf. Figure 4) yielded the best results on test data from Portal 1 regarding the RMSE. All other parameter configurations were adopted from Nguyen and Zhu (2013).

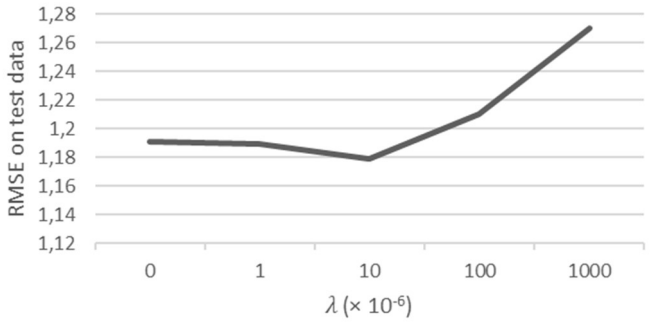

Figure 4. RMSE on the Test Data Depending on the Regularization Penalty Factor λ

Test Procedure and Results

For our evaluation, we split ratings into 50% training data for learning the CBMF model and 50% test data for assessing the prediction accuracy. On the one hand, dividing the data in half at random allowed to obtain a large *test* set (cf. also Nguyen and Zhu 2013), which is important for meaningful results when testing hypotheses. On the other hand, because of the large real-world data sets, 50% *training* data allowed us to learn the CBMF model.

After that, we utilized the recommender system for each pair of item content data sets (with and without increased completeness) to predict ratings and assess the corresponding prediction accuracy. The increase in prediction accuracy assessed separately by Precision, Recall and F1-measure was determined by subtracting the prediction

accuracy based on the baseline content from the prediction accuracy based on the content with increased completeness. As lower RMSE values indicate more accurate predictions, the negative difference was used in this case, accordingly.

A requirement for evaluating Hypotheses H1a/b using Student's t-test is that sample groups should be normally distributed. Because of the large sample size in our evaluation, this requirement is obviously met (Boneau 1960). For evaluating moderator effects in Hypotheses H2-H5, we examined whether the selection of the linear regression model is appropriate or whether non-linear, for instance, quadratic regression models should be preferred (i.e., a curvilinear moderator effect is expected). Therefore, to test for potential non-linear moderator effects, we compared the fitness of the quadratic (non-linear) model and the linear model relying on the frequently discussed and used Bayesian Information Criterion (BIC) for model selection (cf. Schwarz 1978), for which smaller BIC values indicate the preferred model. These tests yielded almost the same BIC values for both models. For instance, for the first Hypothesis H2a the BIC value for the linear model was -5,464 and for the quadratic model -5,446 (e.g., regarding the measure Precision) and for the last Hypothesis H5b the BIC value for the linear model was -155 and for the quadratic model -148. Since the quadratic model did not or hardly improve the BIC values, the linear model was used because of its lower complexity, as suggested by literature (Cohen et al. 2003; MacCallum and Mar 1995).

The moderator variable for Hypotheses H2a/b was operationalized by the number of added or filled up feature value assignments *per item* (cf. Blake and Mangiameli 2011) relative to the number of feature value assignments *per item* in the baseline content data set. For Hypotheses H3a/b, the mean of the aforementioned operationalization across all rated items of *a user* was used as the moderator variable. In a similar way, the moderator variable for Hypothesis H4a was operationalized by the number of added feature value assignments for *a feature* relative to the number of feature value assignments in the baseline content data set. Hypothesis H4b was operationalized by the number of filled up feature value assignment for *a feature* relative to the number of feature value assignments for *this feature* in the baseline data set. The moderator for Hypotheses H5a/b was assessed by the mean cosine distance between the added/filled up features and the baseline features (Mitra et al. 2002; Tabakhi and Moradi 2015). Summing up the above, each operationalization of the moderator variables shares a similar concept as it was determined as the increase in completeness relative to the baseline content.

Furthermore, we used the two standard levels of significance 0.01 (indicated by '**') and 0.05 (indicated by '*') for the tests of all hypotheses (e.g., Shani and Gunawardana 2011).

In the following, we outline the evaluation results. In particular, we present the impact on prediction accuracy for all tests, which means, the values for each measure (RMSE, Precision, Recall and F1-measure), their relative increase

in prediction accuracy and the significance of the t-values in case of H1a/b and the significance of the regression coefficients together with the effect sizes in case of H2-H5. Table 7 shows the results of our evaluation for the first two hypotheses: Hypotheses H1a and H1b can be supported with positive t-values and statistical significance by p-values below 0.01. This means that both adding features and their feature values as well as filling up missing feature values lead to significantly higher prediction accuracy as indicated by each of the evaluation measures in Table 7.

Hypothesis (Origin of Rating Data)	**Compared Data Sets**	**Prediction Accuracy** (RMSE/Precision/ Recall/F1) (Without Increased Completeness)	**Prediction Accuracy** (RMSE/Precision/ Recall/F1) (Increased Completeness)	**Relative Increase in Prediction Accuracy** (RMSE/Precision/ Recall/F1)	**Corresponding t-Values** (*:p-value<0.05; **:p-value<0.01)	**Hypothesis can be supported**
H1a (Portal 1)	P1.1 vs. P1.1&P2	1.57/0.216/ 0.218/0.217	1.18/0.246/ 0.231/0.238	25%/14%/ 6%/10%	164**/63**/ 63**/63**	**Yes** (by all)
H1a (Portal 2)	P2 vs. P1.1&P2	1.29/0.236/ 0.235/0.235	1.20/0.249/ 0.246/0.247	7%/6%/ 5%/5%	17**/5**/ 5**/5**	**Yes** (by all)
H1a (movie-lens)	OldV vs. OldV&NewV	1.67/0.226/ 0.228/0.227	0.95/0.443/ 0.315/0.368	43%/96%/ 38%/62%	413**/185**/ 185**/185**	**Yes** (by all)
H1b (Portal 1)	P1.2 vs. P1.2'	1.60/0.227/ 0.221/0.224	1.04/0.332/ 0.225/0.268	35%/46%/ 2%/20%	269**/112**/ 112**/112**	**Yes** (by all)

Table 7. Results for Hypotheses H1a/b

Hypothesis (Origin of Rating Data)	**Compared Data Sets**	**Interaction Coefficients b_3 of Moderated Regression Model with Dependent Variable RMSE/Precision/ Recall/F1** (*:p-value<0.05; **:p-value<0.01)	**Cohen's f^2 of Moderated Regression Model with Dependent Variable RMSE/Precision/ Recall/F1**	**Hypothesis can be supported**
H2a (Portal 1)	P1.1 vs. P1.1&P2	0.06**/0.02**/ 0.01**/0.01**	0.024/0.014/ 0.002/0.008	**Yes** (by all)
H2a (Portal 2)	P2 vs. P1.1&P2	0.12**/0.02**/ 0.02**/0.02**	0.042/0.001/ 0.001/0.001	**Yes** (by all)
H2a (movielens)	OldV vs. OldV& NewV	0.03**/0.01**/ 0.001**/0.004**	0.032/0.015/ 0.001/0.011	**Yes** (by all)
H2b (Portal 1)	P1.2 vs. P1.2'	0.24**/0.03**/ 0.02**/0.02**	0.436/0.019/ 0.009/0.015	**Yes** (by all)
H3a (Portal 1)	P1.1 vs. P1.1&P2	0.08**/0.02**/ 0.01**/0.01**	0.013/0.002/ 0.001/0.001	**Yes** (by all)
H3a (Portal 2)	P2 vs. P1.1&P2	0.18**/-0.01/ 0.00/-0.01	0.015/-/ -/-	**Only for RMSE measure**
H3a (movielens)	OldV vs. OldV& NewV	0.02**/0.004**/ 0.004**/0.004**	0.018/0.003/ 0.003/0.005	**Yes** (by all)
H3b (Portal 1)	P1.2 vs. P1.2'	0.39**/0.01*/ 0.02**/0.01**	0.094/0.0003/ 0.001/0.001	**Yes** (by all)

Table 8. Results for Hypotheses H2a/b and H3a/b

The results of the hypotheses with regard to items and users are given in Table 8: Hypotheses H2a and H2b can also be supported with statistical significance by p-values below 0.01. This means that for items both the amount of additional features and their feature values and the amount of filled up feature values are positive moderators. In other words, items that obtain a stronger increase in completeness can then be recommended at a significant higher level of accuracy than before (cf. Table 8). For Hypotheses H3a and H3b, focusing on users instead of items, the test results were as follows: Hypothesis H3a in the case of Portal 1 and movielens as well as Hypothesis H3b can be supported with statistical significance by p-values below 0.01 (except for the case of the measure Precision for H3b, where the p-value was between 0.01 and 0.05). The test of Hypothesis H3a in the case of Portal 2 yielded a p-value below 0.01 only for the measure RMSE, but p-values above 0.05 for the measures Precision, Recall and F1-measure. Hence, Hypothesis H3a cannot be supported for all measures in the case of Portal 2. Except from that, Hypothesis H3 can be supported in the case of Portal 1 and movielens with statistical significance at the level 0.05. Therefore, it can be concluded that both the amount of additional features and their feature values and the amount of filled up

Hypothesis (Origin of Rating Data)	**Compared Data Sets**	**Interaction Coefficients b_3 of Moderated Regression Model with Dependent Variable RMSE/Precision/ Recall/F1** (*:p-value<0.05; **:p-value<0.01)	**Cohen's f^2 of Moderated Regression Model with Dependent Variable RMSE/Precision/ Recall/F1**	**Hypothesis can be supported**
H4a (Portal 1)	P1.1 vs. P1.1&P2	0.40**/0.02**/ 0.02**/0.02**	1.221/0.611/ 0.628/0.645	**Yes** (by all)
H4a (Portal 2)	P2 vs. P1.1&P2	0.27**/0.09**/ 0.08**/0.09**	0.363/0.236/ 0.162/0.198	**Yes** (by all)
H4a (movielens)	OldV vs. OldV&NewV	0.70**/0.10**/ 0.04**/0.07**	1.657/0.665/ 0.233/0.575	**Yes** (by all)
H4b (Portal 1)	P1.2 vs. P1.2'	-0.01/0.00/ 0.00/0.00	-/-/ -/-	**No** (by all)
H5a (Portal 1)	P1.1 vs. P1.1&P2	-1.94**/-0.11**/ -0.10**/-0.11**	0.487/0.297/ 0.367/0.338	**No** (by all)
H5a (Portal 2)	P2 vs. P1.1&P2	-0.02**/-0.01**/ -0.01**/-0.01**	0.040/0.051/ 0.038/0.044	**No** (by all)
H5a (movielens)	OldV vs. OldV&NewV	-0.76**/-0.12**/ -0.07**/-0.09**	0.352/0.280/ 0.280/0.367	**No** (by all)
H5b (Portal 1)	P1.2 vs. P1.2'	-0.43*/-0.05/ 0.00/-0.02	0.155/-/ -/-	**No** (by all)

Table 9. Results for Hypotheses H4a/b and H5a/b

feature values each measured per user are also positive moderators of the impact of completeness on prediction accuracy assessed by RMSE and, except H3a (Portal 2), on prediction accuracy assessed by Precision, Recall and

F1-measure. This means that users, whose rated items obtain a stronger increase in completeness, benefit the most and that recommendations for these users are significantly more accurate than before.

The results of Hypotheses H4a/b and H5a/b are given in Table 9. Hypothesis H4a can be supported with statistical significance by p-values below 0.01, whereas Hypothesis H4b cannot be supported indicated by negative coefficients b_3. In other words, only the amount of additional features and their feature values is a positive moderator of the impact on prediction accuracy (H4a), but not the amount of filled up feature values (H4b). Hypotheses H5a/b cannot be supported as indicated by negative coefficients or by p-values above 0.05. This suggests that the diversity for an additional feature or for a filled up feature is not a positive moderator.

Discussion and Implications

In general, the results support the theoretical model serving as foundation of the tested hypotheses, which means, the completeness of item content data has a significant positive impact on the prediction accuracy of recommendations. More precisely, adding features and their feature values (Hypothesis H1a) or filling up missing feature values (Hypothesis H1b) leads to higher prediction accuracy. Besides this general finding, we also examined moderator effects on the impact of completeness on prediction accuracy (Hypotheses H2-H5). Thereby, the results reveal some interesting findings. While the increase in completeness per item and per user are positive moderators of the impact of completeness on prediction accuracy (Hypotheses H2a/b and H3a/b, except for Hypothesis H3a and Portal 2, which will be discussed below), the same cannot always be examined for the increase in completeness per feature. In particular, adding features with a high amount of additional feature values leads to a higher increase in prediction accuracy (Hypothesis H4a). However, filling up missing feature values with a high amount of additional feature values does not lead to a higher increase in prediction accuracy (Hypothesis H4b). In addition, neither adding features (Hypothesis H5a) nor filling up missing values of features (Hypothesis 5b) with a high diversity leads to a higher increase in prediction accuracy, which constitutes a further interesting finding. In the following, we discuss each result in detail.

Both Hypotheses H1a and H1b are supported as indicated by t-values with positive sign and with p-values below 0.01. This means, as illustrated in Table 7, both adding features and their feature values as well as filling up missing feature values led to a considerable increase in prediction accuracy. After increasing completeness, the RMSE was between 7% and 43% lower than the RMSE before increasing completeness (corresponding to absolute decreases of RMSE between 0.09 and 0.72). Precision was between 6% and 96% higher, Recall was between 2% and 38% higher and F1-measure was between 5% and 62% higher. By a detailed consideration of the results for Hypotheses H1a/b,

two interesting observations can be made. First, the relative increase in prediction accuracy is lower for H1a in the case of Portal 2 compared to all other cases of H1a. This may be due to the fact that the additional features only constitute less than 40% of all features of the item content data set with increased completeness in case of Portal 2 (7 of 19 features). In the other cases of H1a, the additional features constitute at minimum 60% of the features of the data set with increased completeness (12 of 19 features or 5 of 6 features). Second, the increase in prediction accuracy measured by RMSE and Precision is in almost all cases (considerably) higher than measured by Recall and F1-measure. In contrast to the discrete nature of the measures Precision, Recall and the F1-measure, the higher increase in prediction accuracy measured by RMSE may be reasoned by the fact that RMSE uses the predicted ratings as determined by the recommender system (i.e., as a continuous variable). Therefore, the errors between predicted and actual ratings are assessed by an interval-scaled difference. To analyze the high increases in Precision, we examined the results of H1a (movielens) in more detail, which shows the highest increase of Precision (+96%). Here, we found that, on the one hand, the *decreases* in the number of incorrect predictions (i.e., false positives) was the largest for the rating levels 1 star (-96%), 2 stars (-76%) and 5 stars (-60%). On the other hand, the largest *increases* in correct predictions (true positives) was achieved for the ratings levels 3 stars (+31%) and 4 stars (+207%). This means that by increasing completeness the used recommender system was less likely to incorrectly predict "extreme" ratings (i.e., very high or very low ratings) while mostly improving the correct prediction of "mainstream" ratings (the mean overall rating is 3.6). Hence, the Precision of most classes achieved a much higher increase than the Recall or F1-measure. In total, the results of Hypotheses H1a/b show that recommendations based on item content data sets with increased completeness are more accurate, which is valuable for achieving a high user satisfaction (Koren et al. 2009; Ricci et al. 2015). At this point, we want to emphasize that the increase in prediction accuracy is provided only by increasing data quality and not by enhancing the recommender algorithm. Nowadays, the aim of numerous works in the research field of recommender systems is to develop very sophisticated recommender algorithms in order to increase prediction accuracy (partly to a small extent). One seminal example is the winning solution of the Netflix Grand Prize, which decreased the RMSE by 10% through a very elaborate and complex enhancement and combination of multiple recommender algorithms (Koren 2009). Instead, our results show that devoting more importance to maintaining high data quality for recommender systems is also highly promising and may inspire further research.

For Hypotheses H2-H5 we focus on the coefficients b_3 regarding the moderated regression (cf. Equation (5)) as well as the corresponding effect sizes indicated by Cohen's f^2 (cf. Equation (7)). Here, in general, the absolute values of the coefficients b_3 are consistently higher when evaluating the RMSE compared to the other measures. This is due to the higher values for the RMSE as seen in Table 7, where the values for RMSE range from 0.95 to 1.67 while

Precision, Recall and F1-measure take values between 0.216 and 0.443. Considering the results for Hypothesis H2b, for instance, the coefficient for the RMSE signifies that the RMSE based on increased completeness is lowered by 0.24 when the moderator variable is increased by one. In the same setting, the Precision would increase by only 0.03.

The evaluation results support Hypotheses H2a/b. As illustrated in Table 8, all coefficients b_3 of our evaluation were positive (ranging from 0.001 to 0.24) and significant (p-value<0.01). This finding shows that the amount of additional features and their feature values and the amount of filled up feature values *per item* has a significant moderator effect. The effect size indicated by Cohen's f^2 ranges from 0.001 to 0.436 (cf. Section "Model for Hypotheses H2-H5" for the interpretation of Cohen's f^2). By a detailed consideration of the results for Hypotheses H2a/b, three observations can be made. First, the evaluation measure RMSE showed the largest effect sizes. This is in accordance with the finding discussed above that prediction accuracy measured by RMSE shows the highest increase in general due to its continuous nature. Second, the effect sizes for Precision, Recall and F1-measure, especially for H2a (Portal 2), are small. This may be reasoned by similar arguments as the first observation and by the fact, that the additional features only constitute less than 40% of all features of the item content data set, as discussed above for H1a (Portal 2). Third, the effect size for Hypothesis H2b is relatively high. An analysis of the data indicated that items, which have many missing feature values, receive highly incorrect rating predictions based on the data set without increased completeness (i.e., the baseline prediction accuracy is low). Therefore, these items benefit considerably from increased completeness in terms of prediction accuracy. The findings above should encourage web portals and business owners to increase and maintain the completeness of item content data. In addition, the results of Hypotheses H2a/b can be used to balance the cost and benefit of data quality improvement measures, a topic discussed in recent literature (Heinrich et al. 2018a). For instance, only items (e.g., products offered by a web portal) with a higher profit margin can be extended with additional content in a selective manner, avoiding a potentially expensive large-scale extension of the whole data set. This opens up an effective option to manage the item content data in an affordable manner, which can be a crucial factor for web portals.

Hypotheses H3a/b can be also supported except in the case of Portal 2 regarding the measures Precision, Recall and the F1-measure. In all other cases of H3a/b, our evaluation yields significant coefficients b_3 ranging from 0.004 to 0.39. This means that the amount of additional features and their feature values and the amount of filled up feature values *per user* show moderator effects. The effect size indicated by Cohen's f^2 ranges from 0.0003 to 0.094. By a detailed consideration of the results for Hypotheses H3a/b, two interesting observations can be made. First, similarly to the discussions above, RMSE shows the largest effect sizes. Second, in the case of H3a (Portal 2) the p-values of the coefficient b_3 were above the significance level of 0.05 for the measures Precision, Recall and F1-measure. This

may be reasoned by the lower additional item content (7 of 19 features) as well as the lower number of users (505 users) in this particular evaluation. Thus, according to the results of H3 users with a stronger increase in the amount of additional features and their feature values or in the amount of filled up feature values are suggested to have a significantly higher increase in prediction accuracy. This means that web portals – similar to the discussion above – can manage and increase the prediction accuracy for specific users (e.g., users with low versus high sales volumes) by extending the content of items, which have been rated by these users or which may be interesting and recommended for them in the future. In addition, another promising option would be to give providers as well as users, which mainly rate items with a lower number of available features, an incentive to provide additional data for these items. In return, the user community would benefit in this way from more appropriate item recommendations.

The results of Hypotheses H4a/b and H5a/b indicate that the *amount* and *diversity* of additional item content does *in general* not moderate the increase in prediction accuracy as intuition might suggest. Although Hypothesis H4a can be supported by our evaluation with relatively high moderator effects indicated by Cohen's f^2 ranging from 0.162 to 1.657 and with positive significant coefficients b_3 (ranging from 0.02 to 0.70), Hypothesis H4b cannot be supported. This means that portals aiming to extend item content data should primarily focus on (selected) additional features with a high amount of feature values, but filling up features with a high amount of additional feature values does not lead to a higher increase in prediction accuracy in general. At first sight, this result is counterintuitive, as one would have expected that more filled up feature values would lead to a higher increase in prediction accuracy. A reason why filling up individual features with a high amount of missing values does not result in a higher increase in prediction accuracy – indicated by p-values above 0.05 of the coefficients b_3 for all four evaluation measures – could be that the additional content was inferred by a deficient imputation method. However, this can be rebutted as a significant increase in prediction accuracy was achieved in H1b, which would be also caused by the inferred feature values and thus by the chosen imputation technique. Instead, it is necessary to consider the importance of features in this context. For example, the feature *Special Needs* with values such as *Dog Allowed* and *Good For Dancing* has more missing feature values (i.e., less available feature values) than the feature *Parking Information* with values such as *Bike Parking* and *Private Parking Lot*. Therefore, filling up missing values for *Special Needs* leads to a higher increase in completeness compared to *Parking Information*. However, as transportation (e.g., by bike, car or subway) is an important aspect for restaurant visitors in New York City, features such as *Parking Information* seem to be more important for the majority of users (and thus, may be better maintained by those users) than features such as *Special Needs*. In our evaluation, this importance is indicated by a higher increase in prediction accuracy when filling up feature values, for instance, for the feature *Parking Information* compared to filling up the feature *Special Needs*. This shows that the result of H4b may be caused by

important features having potentially less missing data values in the baseline data set. The results regarding Hypothesis H4a can be reasoned in a similar way. Compared to all other hypotheses, effect sizes regarding Hypothesis H4a are the largest. Here, an analysis of the data of H4a (Portal 1) shows that adding features with a high amount of feature value assignments such as *Special Services* yield a high increase in prediction accuracy. This is reasonable, since the feature *Special Services* has the feature values *Cheap Eats*, *Delivery* and *Take Out* and therefore, *Special Services* seems to constitute an important feature for the user ratings for restaurants in general. This further indicates that important features for users are those features with a high amount of available feature value assignments with regard to Hypothesis H4a. Therefore, it is reasonable, that the effect sizes for the moderator in H4a are the largest. Overall, the results do not indicate that the amount of additional feature values by itself is a positive moderator, but a high amount of available feature value assignments in a data set may be an indicator for the importance of features and its impact on prediction accuracy (cf. H4a).

Hypotheses H5a/b cannot be supported as indicated by coefficients b_3 with negative sign or with p-values above the 0.05 level of significance. This means that a higher diversity of added or filled up features does not yield a higher increase in prediction accuracy. In general, adding a feature with exactly the same feature value assignments as an existing feature to the data set should not yield any increase in prediction accuracy, as stated by the literature (Mitra et al. 2002; Tabakhi and Moradi 2015). Hence, the increase in prediction accuracy caused by adding features to a data set is expected to decrease with the similarity of these additional features to the existing features. Therefore, we would have anticipated that adding and filling up features with a high diversity would enable the recommender system to differentiate items in more details, thus leading to more accurate recommendations to users. However, an analysis shows that even features with a high diversity can be of low importance to users and thus, result in a low increase in prediction accuracy. For example, the additional feature *Production Company* with feature values such as *Paramount Pictures* or *Twentieth Century Fox* brings high diversity to the baseline feature *Genre*, as indicated by a mean cosine distance of 0.96 between the features *Production Company* and *Genre*. Nevertheless, adding only this feature has low impact on the increase in prediction accuracy (e.g., the RMSE decreased only by 0.002). This seems reasonable, as production companies produce diverse movies with different actors, directors and of different genres and therefore, the feature *Production Company* is usually of low importance for the majority of users. This underlines that features exist which have diverse feature value assignments, but their importance is low for users. Contrary to works such as (Mitra et al. 2002; Tabakhi and Moradi 2015), which propose to sort out features with high similarity (i.e., low diversity), this shows that the diversity or similarity of features may only be a subordinate factor for the impact of completeness on the prediction accuracy. In total, the increase in completeness by the

amount of additional feature values (H4) as well as by the diversity of added/filled up features (H5) does not constitute a positive moderator of the impact of completeness on prediction accuracy.

Based on these findings and the above discussion, the contribution of our work to the existing body of knowledge can be outlined. Blake and Mangiameli (2011), Feldman et al. (2018) and Woodall et al. (2015) proposed and substantiated that completeness – in the sense of the amount of available feature values – has a significant impact on evaluation criteria such as decision quality of specific considered decision support systems. Complementary to these works, our results show that not only a higher amount of available feature values, but also adding new features to the feature set can have a significant impact on evaluation criteria of decision support systems and in particular recommender systems. Furthermore, so far, the impact of data quality was validated for different evaluation criteria. The works of Bharati and Chaudhury (2004) and Ge (2009) supported the impact on the evaluation criteria decision-making satisfaction and decision quality. Blake and Mangiameli (2011), Feldman et al. (2018) and Woodall et al. (2015) demonstrated the impact on data mining outcome. Supplementing these findings, our work is the first to analyze the impact of data quality – in particular completeness – on the evaluation criterion prediction accuracy. Moreover, our results show that the impact of data quality can be significantly influenced by moderators. While our findings support the so far not examined statement that the impact on prediction accuracy is moderated by the increase in completeness per item and per user, they show that the amount of additional feature values is not a positive moderator in this regard. Moreover, our findings do not support the intuitive concept that the diversity of features is a positive moderator of the impact of completeness on prediction accuracy.

Following this discussion, notable implications can be concluded for applications in practice. Expanding the discussion above, it is crucial for business owners to provide a large(r) number of features for their businesses and to check whether additional important features are available. The resulting increase in completeness leads to more accurate recommendations of these businesses, which better fit the users' preferences. Similarly, the acquisition of additional data is highly advantageous for web portals. It allows improved recommendations and enhances the efficacy of the web portal. Moreover, our findings should encourage meta portals, which already make use of data from different web sources, to further collect additional features and feature values and, in this way, to provide high quality recommendations. Currently, many meta portals (such as *trivago.com*) mainly focus on the integration of user ratings and reviews from different sources and mostly ignore the impact of an extended item content data set. By recommending items based on data with increased completeness, meta portals can exploit a much higher potential of making high quality product recommendations for customers. In case of limitations in acquiring additional features or feature values, it is important to focus on important additional features, which may be

indicated by a high amount of available feature values. In contrast, a high diversity of additional features is not required.

Conclusions, Limitations and Directions for Future Work

We investigate the impact of the data quality dimension completeness of item content data on prediction accuracy. Based on a theoretical model derived from literature, hypotheses are formulated and substantiated. These hypotheses focus on the impact of adding features and filling up missing feature values on the prediction accuracy of recommendations, which was assessed by the measures RMSE, Precision, Recall and the F1-measure. The hypotheses are evaluated on two real-world data sets, one from the domain of restaurants and another one from the domain of movies. Our results yield that rating predictions are significantly more accurate when more features and feature values are available. Moreover, this impact of completeness on the increase in prediction accuracy is moderated by the amount of additional features and their feature values or the amount of filled up feature values per items and per users. In contrast, this statement does not hold for features. While adding features with a high amount of feature values leads to a higher increase in prediction accuracy, filling up a high amount of feature values or adding features to the existing content with a high diversity does not lead to a higher increase in prediction accuracy. Here, our results suggest that the importance of features to users is an essential factor for the increase in prediction accuracy. Our findings are not only valuable from a scientific perspective but also in practice for business owners as well as for web portals and meta portals.

Our work also has some limitations, which could be starting points for future research. In this paper, we increased completeness by adding features from other web portals as well as by imputing missing feature values. Nevertheless, other approaches to increase completeness are possible. For example, a feature set could be extended with features based on user-generated item tags as proposed by Zhang et al. (2010). Similarly, feature values could be filled up by analyzing additional textual data using text mining to extract non-available feature values (Ghani et al. 2006). Another limitation are the costs of data preparation and computation caused by adding features and their feature values or by filling up missing feature values. In our evaluation settings, the necessary additional time and costs are reasonable: For example, the computation time of CBMF for training and evaluating the model for Hypothesis H1a (Portal 1) was raised from 285 seconds to 488 seconds for all users/items, for Hypothesis H1a (Portal 2) from 10 seconds to 24 seconds. However, these costs might indeed be relevant for applications with a vast amount of additional item content data. Furthermore, it would be highly interesting to test the impact of other data quality dimensions such as currency on recommendation quality. Additionally, in this paper we focus on different metrics for prediction accuracy as the most important quality measures for recommender systems (Shani and Gunawardana

2011). However, as the goals of a recommender system can be very diverse (e.g., introducing customers to the full product spectrum) further metrics can be of particular interest for other application scenarios (Jannach et al. 2016). Thus, further research on the impact of data quality assessed by other quality measures such as coverage, serendipity or scalability (Herlocker et al. 2004; Shani and Gunawardana 2011) would also be relevant. Finally, in the future, tests similar to ours could also be conducted using data sets from further domains, such as recommendations for music songs.

References

Abel, Fabian; Herder, Eelco; Houben, Geert-Jan; Henze, Nicola; Krause, Daniel (2013): Cross-system user modeling and personalization on the Social Web. In *User Modeling and User-Adapted Interaction* 23 (2-3), pp. 169–209. DOI: 10.1007/s11257-012-9131-2.

Adomavicius, G.; Tuzhilin, A. (2005): Toward the next generation of recommender systems. A survey of the state-of-the-art and possible extensions. In *IEEE Trans. Knowl. Data Eng.* 17 (6), pp. 734–749. DOI: 10.1109/TKDE.2005.99.

Adomavicius, Gediminas; Zhang, Jingjing (2012): Impact of data characteristics on recommender systems performance. In *ACM Trans. Manage. Inf. Syst.* 3 (1), pp. 1–17. DOI: 10.1145/2151163.2151166.

Adomavicius, Gediminas; Zhang, Jingjing (2016): Classification, Ranking, and Top-K Stability of Recommendation Algorithms. In *INFORMS Journal on Computing* 28 (1), pp. 129–147. DOI: 10.1287/ijoc.2015.0662.

Aggarwal, Charu C. (2014): Data Classification: Chapman and Hall/CRC.

Aguinis, Herman; Beaty, James C.; Boik, Robert J.; Pierce, Charles A. (2005): Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. In *The Journal of applied psychology* 90 (1), pp. 94–107. DOI: 10.1037/0021-9010.90.1.94.

Amatriain, Xavier; Pujol, Josep M.; Tintarev, Nava; Oliver, Nuria (2009): Rate it again. In Lawrence Bergman, Alex Tuzhilin, Robin Burke, Alexander Felfernig, Lars Schmidt-Thieme (Eds.): Proceedings of the third ACM conference on Recommender systems. New York, New York, USA. ACM Special Interest Group on Computer-Human Interaction. New York, NY: ACM, pp. 173–180.

Ballou, Donald P.; Pazer, Harold L. (1985): Modeling data and process quality in multi-input, multi-output information systems. In *Management Science* 31 (2), pp. 150–162.

Basaran, Daniel; Ntoutsi, Eirini; Zimek, Arthur (2017): Redundancies in Data and their Effect on the Evaluation of Recommendation Systems: A Case Study on the Amazon Reviews Datasets. In Nitesh Chawla, Wei Wang (Eds.): Proceedings of the 2017 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, pp. 390–398.

Batini, Carlo; Cappiello, Cinzia; Francalanci, Chiara; Maurino, Andrea (2009): Methodologies for data quality assessment and improvement. In *ACM Comput. Surv.* 41 (3), pp. 1–52. DOI: 10.1145/1541880.1541883.

Batini, Carlo; Scannapieco, Monica (2016): Data and Information Quality. Cham: Springer International Publishing.

Bell, Robert M.; Koren, Yehuda; Volinsky, Chris (2007): The BellKor solution to the Netflix prize.

Berkovsky, Shlomo; Kuflik, Tsvi; Ricci, Francesco (2012): The impact of data obfuscation on the accuracy of collaborative filtering. In *Expert Systems with Applications* 39 (5), pp. 5033–5042. DOI: 10.1016/j.eswa.2011.11.037.

Bharati, Pratyush; Chaudhury, Abhijit (2004): An empirical investigation of decision-making satisfaction in web-based decision support systems. In *Decision Support Systems* 37 (2), pp. 187–197. DOI: 10.1016/S0167-9236(03)00006-X.

Blake, Roger; Mangiameli, Paul (2011): The Effects and Interactions of Data Quality and Problem Complexity on Classification. In *J. Data and Information Quality* 2 (2), pp. 1–28. DOI: 10.1145/1891879.1891881.

Bobadilla, Jesús; Ortega, Fernando; Hernando, Antonio; Gutiérrez, Abraham (2013): Recommender systems survey. In *Knowledge-Based Systems* 46, pp. 109–132.

Boneau, C. Alan (1960): The effects of violations of assumptions underlying the t test. In *Psychological Bulletin* 57 (1), pp. 49–64. DOI: 10.1037/h0041412.

Bostandjiev, Svetlin; O'Donovan, John; Höllerer, Tobias (2012): TasteWeights: a visual interactive hybrid recommender system. In Pádraig Cunningham, Neil Hurley, Ido Guy, Sarabjot Singh Anand (Eds.): Proceedings of the sixth ACM conference on Recommender systems. Dublin, Ireland. ACM Special Interest Group on

Electronic Commerce; ACM Special Interest Group on Knowledge Discovery in Data; ACM Special Interest Group on Artificial Intelligence; ACM Special Interest Group on Computer-Human Interaction; ACM Special Interest Group on Hypertext, Hypermedia, and Web; ACM Special Interest Group on Information Retrieval. New York, NY: ACM, pp. 35–42.

Burke, Robin; Ramezani, Maryam (2011): Matching Recommendation Technologies and Domains. In Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor (Eds.): Recommender Systems Handbook. Boston, MA: Springer US, pp. 367–386.

Christen, Peter (2012): Data matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin, Heidelberg: Springer Berlin Heidelberg.

Cohen, Jacob (1988): Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Erlbaum. Available online at http://gbv.eblib.com/patron/FullRecord.aspx?p=1192162.

Cohen, Jacob; Cohen, Patricia; West, Stephen G.; Aiken, Leona S. (2003): Applied multiple regression/correlation analysis for the behavioral sciences. third edition. New York, London, Mahwah, NJ: Routledge Taylor & Francis Group. Available online at http://www.loc.gov/catdir/enhancements/fy0634/2002072068-d.html.

Cunha, Tiago; Soares, Carlos; Carvalho, André C. P. L. F. de (2016): Selecting Collaborative Filtering Algorithms Using Metalearning. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, Jilles Vreeken (Eds.): Machine Learning and Knowledge Discovery in Databases. European Conference, Ecml Pkdd 2016, Riva Del Garda, Italy, September 19-23, 2016, Proceedings, vol. 9852. Cham: Springer-Verlag New York Inc (LNCS Sublibrary: SL7 - Artificial Intelligence, 9851-9853), pp. 393–409.

Dawson, Jeremy F. (2014): Moderation in Management Research: What, Why, When, and How. In *J Bus Psychol* 29 (1), pp. 1–19. DOI: 10.1007/s10869-013-9308-7.

Doerfel, Stephan; Jäschke, Robert; Stumme, Gerd (2016): The Role of Cores in Recommender Benchmarking for Social Bookmarking Systems. In *ACM Trans. Intell. Syst. Technol.* 7 (3), pp. 1–33. DOI: 10.1145/2700485.

Ekstrand, Michael; Riedl, John (2012): When recommenders fail. In Pádraig Cunningham (Ed.): Proceedings of the sixth ACM conference on Recommender systems. the sixth ACM conference. Dublin, Ireland, 9/9/2012 - 9/13/2012. New York, NY: ACM (ACM Digital Library), p. 233.

Enders, Craig K. (2010): Applied missing data analysis. New York: Guilford Press (Methodology in the social sciences). Available online at http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10389908.

Feldman, Michael; Even, Adir; Parmet, Yisrael (2018): A methodology for quantifying the effect of missing data on decision quality in classification problems. In *Communications in Statistics–Theory and Methods* 47 (11), pp. 2643–2663.

Felfernig, Alexander; Friedrich, Gerhard; Schmidt-Thieme, Lars (2007): Recommender systems. In *IEEE Intelligent Systems* 22 (3).

Forbes, Peter; Zhu, Mu (2011): Content-boosted matrix factorization for recommender systems. In Bamshad Mobasher, Robin Burke, Dietmar Jannach, Gediminas Adomavicius (Eds.): Proceedings of the fifth ACM conference on Recommender systems. Proceedings of the fifth ACM conference on Recommender systems. Chicago, Illinois, USA. New York, NY: ACM, pp. 261–264.

Fortes, Reinaldo Silva; Freitas, Alan R. R. de; Gonçalves, Marcos André (2017): A Multicriteria Evaluation of Hybrid Recommender Systems: On the Usefulness of Input Data Characteristics.

Ge, Mouzhi (2009): Information quality assessment and effects on inventory decision-making. Doctoral dissertation. Dublin City University, Dublin City University.

Geuens, Stijn; Coussement, Kristof; Bock, Koen W. de (2018): A framework for configuring collaborative filtering-based recommendations derived from purchase data. In *European Journal of Operational Research* 265 (1), pp. 208–218. DOI: 10.1016/j.ejor.2017.07.005.

Ghani, Rayid; Probst, Katharina; Liu, Yan; Krema, Marko; Fano, Andrew (2006): Text mining for product attribute extraction. In *ACM SIGKDD Explorations Newsletter* 8 (1), pp. 41–48. DOI: 10.1145/1147234.1147241.

Gignac, Gilles E.; Szodorai, Eva T. (2016): Effect size guidelines for individual differences researchers. In *Personality and Individual Differences* 102, pp. 74–78. DOI: 10.1016/j.paid.2016.06.069.

Gomez-Uribe, Carlos A.; Hunt, Neil (2016): The Netflix recommender system: Algorithms, business value, and innovation. In *ACM Transactions on Management Information Systems (TMIS)* 6 (4, Article 13).

Grčar, Miha; Mladenič, Dunja; Fortuna, Blaž; Grobelnik, Marko (2006): Data Sparsity Issues in the Collaborative Filtering Framework. In Olfa Nasraoui (Ed.): Advances in web mining and web usage analysis. 7th International Workshop on Knowledge Discovery on the Web, WebKDD 2005 : Chicago, IL, USA, August 21, 2005 : revised papers, vol. 4198. Berlin: Springer (Lecture Notes in Computer Science, 4198), pp. 58–76.

Griffith, Josephine; O'Riordan, Colm; Sorensen, Humphrey (2012): Investigations into user rating information and predictive accuracy in a collaborative filtering domain. In Sascha Ossowski, Paola Lecca (Eds.): Proceedings of the 27th annual ACM symposium on applied computing 2012. Symposium on Applied Computing : Riva del

Garda, Trento, Italy, March 26-30, 2012. the 27th Annual ACM Symposium. Trento, Italy, 3/26/2012 - 3/30/2012. New York, N.Y.: ACM Press; Association for Computing Machinery, p. 937.

Gunawardana, Asela; Shani, Guy (2015): Evaluating Recommender Systems. In Francesco Ricci, Lior Rokach, Bracha Shapira (Eds.): Recommender Systems Handbook, vol. 12. Boston, MA: Springer US, pp. 265–308.

Harper, F. Maxwell; Konstan, Joseph A. (2015): The MovieLens Datasets. In *ACM Trans. Interact. Intell. Syst.* 5 (4), pp. 1–19. DOI: 10.1145/2827872.

Hayes, Andrew F. (2013): Introduction to mediation, moderation, and conditional process analysis. A regression-based approach. New York, NY: Guilford Press (Methodology in the social sciences). Available online at http://lib.myilibrary.com/detail.asp?id=480011.

Heinrich, Bernd; Hristova, Diana (2016): A quantitative approach for modelling the influence of currency of information on decision-making under uncertainty. In *Journal of Decision Systems* 25 (1), pp. 16–41. DOI: 10.1080/12460125.2015.1080494.

Heinrich, Bernd; Hristova, Diana; Klier, Mathias; Schiller, Alexander; Szubartowicz, Michael (2018a): Requirements for Data Quality Metrics. In *J. Data and Information Quality* 9 (2), pp. 1–32. DOI: 10.1145/3148238.

Heinrich, Bernd; Klier, Mathias; Schiller, Alexander; Wagner, Gerit (2018b): Assessing data quality – A probability-based metric for semantic consistency. In *Decision Support Systems* 110, pp. 95–106. DOI: 10.1016/j.dss.2018.03.011.

Helm, Roland; Mark, Antje (2012): Analysis and evaluation of moderator effects in regression models: state of art, alternatives and empirical example. In *Rev Manag Sci* 6 (4), pp. 307–332. DOI: 10.1007/s11846-010-0057-y.

Herlocker, Jonathan L.; Konstan, Joseph A.; Terveen, Loren G.; Riedl, John T. (2004): Evaluating collaborative filtering recommender systems. In *ACM Transactions on Information Systems (TOIS)* 22 (1), pp. 5–53.

Huang, Zan; Zeng, Daniel D. (2005): Why Does Collaborative Filtering Work? Recommendation Model Validation and Selection By Analyzing Bipartite Random Graphs. In *SSRN Journal*. DOI: 10.2139/ssrn.894029.

Jannach, Dietmar; Resnick, Paul; Tuzhilin, Alexander; Zanker, Markus (2016): Recommender Systems - Beyond Matrix Completion. In *Commun. ACM* 59 (11), pp. 94–102. DOI: 10.1145/2891406.

Karatzoglou, Alexandros; Hidasi, Balázs (2017): Deep Learning for Recommender Systems. In Paolo Cremonesi, Francesco Ricci, Shlomo Berkovsky, Alexander Tuzhilin (Eds.): Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17. the Eleventh ACM Conference. Como, Italy, 27.08.2017 - 31.08.2017. New York, New York, USA: ACM Press, pp. 396–397.

Kayaalp, Mehmet; Özyer, Tansel; Özyer, Sibel Tariyan (2009): A Collaborative and Content Based Event Recommendation System Integrated with Data Collection Scrapers and Services at a Social Networking Site. In Nasrullah Memon (Ed.): International Conference on Advances in Social Networks Analysis and Mining, 2009. Piscataway, NJ: IEEE, pp. 113–118.

Kim, Donghyun; Park, Chanyoung; Oh, Jinoh; Lee, Sungyoung; Yu, Hwanjo (2016): Convolutional Matrix Factorization for Document Context-Aware Recommendation. In Shilad Sen, Werner Geyer, Jill Freyne, Pablo Castells (Eds.): Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16. the 10th ACM Conference. Boston, Massachusetts, USA, 15.09.2016 - 19.09.2016. New York, New York, USA: ACM Press, pp. 233–240.

Konstan, Joseph A.; Riedl, John (2012): Recommender systems. From algorithms to user experience. In *User Model User-Adap Inter* 22 (1-2), pp. 101–123. DOI: 10.1007/s11257-011-9112-x.

Koren, Yehuda (2009): The bellkor solution to the netflix grand prize. In *Netflix prize documentation* 81, pp. 1–10.

Koren, Yehuda; Bell, Robert; Volinsky, Chris (2009): Matrix Factorization Techniques for Recommender Systems. In *Computer* 42 (8), pp. 30–37. DOI: 10.1109/MC.2009.263.

Lathia, Neal; Amatriain, Xavier; Pujol, Josep M. (2009): Collaborative filtering with adaptive information sources. In Sarabjot Singh Anand, Bamshad Mobasher, Alfred Kobsa, Dietmar Jannach (Eds.): Proceedings of the 7th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems (ITWP'09). Intelligent Techniques for Web Personalization & Recommender Systems -- ITWP'09. Pasadena, California, USA, July 11-17. CEUR-WS. org (CEUR Workshop Proceedings (CEUR-WS.org), 528), pp. 81–86.

Lee, Yang W.; Strong, Diane M.; Kahn, Beverly K.; Wang, Richard Y. (2002): AIMQ: a methodology for information quality assessment. In *Information & Management* 40 (2), pp. 133–146. DOI: 10.1016/S0378-7206(02)00043-5.

Levi, Asher; Mokryn, Osnat; Diot, Christophe; Taft, Nina (2012): Finding a needle in a haystack of reviews. cold start context-based hotel recommender system. In Pádraig Cunningham, Neil Hurley, Ido Guy, Sarabjot Singh Anand (Eds.): Proceedings of the sixth ACM conference on Recommender systems. Dublin, Ireland. ACM Special Interest Group on Electronic Commerce; ACM Special Interest Group on Knowledge Discovery in Data; ACM Special Interest Group on Artificial Intelligence; ACM Special Interest Group on Computer-Human

Interaction; ACM Special Interest Group on Hypertext, Hypermedia, and Web; ACM Special Interest Group on Information Retrieval. New York, NY: ACM, pp. 115–122.

Levy, Yair; Ellis, Timothy J. (2006): A systems approach to conduct an effective literature review in support of information systems research. In *Informing Science* 9, pp. 181–212.

Li, Seth Siyuan; Karahanna, Elena (2015): Online recommendation systems in a B2C E-commerce context: a review and future directions. In *Journal of the Association for Information Systems* 16 (2), pp. 72–107.

Lops, Pasquale; Gemmis, Marco de; Semeraro, Giovanni (2011): Content-based recommender systems. State of the art and trends. In : Recommender systems handbook: Springer, pp. 73–105.

Lu, Jie; Wu, Dianshuang; Mao, Mingsong; Wang, Wei; Zhang, Guangquan (2015): Recommender system application developments: A survey. In *Decision Support Systems* 74, pp. 12–32. DOI: 10.1016/j.dss.2015.03.008.

MacCallum, Robert C.; Mar, Corinne M. (1995): Distinguishing between moderator and quadratic effects in multiple regression. In *Psychological Bulletin* 118 (3), pp. 405–421. DOI: 10.1037/0033-2909.118.3.405.

Matuszyk, Pawel; Spiliopoulou, Myra (2014): Predicting the Performance of Collaborative Filtering Algorithms. In Rajendra Akerkar, Nick Bassiliades, John Davies, Vadim Ermolayev (Eds.): WIMS '14 : 4th International Conference on Web Intelligence, Mining and Semantics. the 4th International Conference. Thessaloniki, Greece, 6/2/2014 - 6/4/2014. New York, New York, USA: ACM Press, pp. 1–6.

Mitra, P.; Murthy, C. A.; Pal, S. K. (2002): Unsupervised feature selection using feature similarity. In *IEEE Trans. Pattern Anal. Machine Intell.* 24 (3), pp. 301–312. DOI: 10.1109/34.990133.

Nguyen, Jennifer; Zhu, Mu (2013): Content-boosted matrix factorization techniques for recommender systems. In *Statistical Analy Data Mining* 6 (4), pp. 286–301. DOI: 10.1002/sam.11184.

Ning, Xia; Desrosiers, Christian; Karypis, George (2015): A comprehensive survey of neighborhood-based recommendation methods. In : Recommender systems handbook: Springer, pp. 37–76.

Ning, Yue; Shi, Yue; Hong, Liangjie; Rangwala, Huzefa; Ramakrishnan, Naren (2017): A Gradient-based Adaptive Learning Framework for Efficient Personal Recommendation. In Paolo Cremonesi, Francesco Ricci, Shlomo Berkovsky, Alexander Tuzhilin (Eds.): Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17. the Eleventh ACM Conference. Como, Italy, 27.08.2017 - 31.08.2017. New York, New York, USA: ACM Press, pp. 23–31.

Olteanu, Alexandra; Kermarrec, Anne-Marie; Aberer, Karl (2014): Comparing the Predictive Capability of Social and Interest Affinity for Recommendations. In Boualem Benatallah, Azer Bestavros, Yannis Manolopoulos, Athena Vakali, Yanchun Zhang (Eds.): Web information systems engineering - WISE 2014. 15th International Conference, Thessaloniki, Greece, October 12-14, 2014 : proceedings, vol. 8786. Cham: Springer (LNCS sublibrary. SL 3, Information systems and application, incl. Internet/Web and HCI, 8786-8787), pp. 276–292.

Ozsoy, Makbule Gulcin; Polat, Faruk; Alhajj, Reda (2015): Modeling Individuals and Making Recommendations Using Multiple Social Networks. In Jian Pei, Fabrizio Silvestri, Jie Tang (Eds.): Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway, NJ, New York, NY: IEEE; ACM, pp. 1184–1191.

Pazzani, Michael J.; Billsus, Daniel (2007): Content-based recommendation systems. In : The adaptive web: Springer, pp. 325–341.

Pessemier, Toon de; Dooms, Simon; Deryckere, Tom; Martens, Luc (2010): Time dependency of data quality for collaborative filtering algorithms. In Xavier Amatriain, Marc Torrens, Paul Resnick, Markus Zanker (Eds.): Proceedings of the fourth ACM conference on Recommender systems. Barcelona, Spain. ACM Special Interest Group on Knowledge Discovery in Data; ACM Special Interest Group on Electronic Commerce; ACM Special Interest Group on Artificial Intelligence; ACM Special Interest Group on Computer-Human Interaction; ACM Special Interest Group on Information Retrieval; ACM Special Interest Group on Hypertext, Hypermedia, and Web. New York, NY: ACM, pp. 281–284.

Picault, Jérome; Ribiere, Myriam; Bonnefoy, David; Mercer, Kevin (2011): How to get the Recommender out of the Lab? In Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor (Eds.): Recommender Systems Handbook. Boston, MA: Springer US, pp. 333–365.

Pipino, Leo L.; Lee, Yang W.; Wang, Richard Y. (2002): Data quality assessment. In *Commun. ACM* 45 (4), pp. 211–218. DOI: 10.1145/505248.506010.

Porcel, Carlos; Herrera-Viedma, Enrique (2010): Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries. In *Knowledge-Based Systems* 23 (1), pp. 32–39.

Power, Daniel J.; Sharda, Ramesh; Burstein, Frada (2015): Decision support systems: John Wiley & Sons, Ltd.

Redman, Thomas C. (1996): Data quality for the information age. Boston, MA: Artech House (The Artech House computer science library).

Ricci, Francesco; Rokach, Lior; Shapira, Bracha (2015): Recommender Systems: Introduction and Challenges. In Francesco Ricci, Lior Rokach, Bracha Shapira (Eds.): Recommender Systems Handbook. Boston, MA: Springer US, pp. 1–34.

Ricci, Francesco; Rokach, Lior; Shapira, Bracha; Kantor, Paul B. (Eds.) (2011): Recommender Systems Handbook. Boston, MA: Springer US.

Sar Shalom, Oren; Berkovsky, Shlomo; Ronen, Royi; Ziklik, Elad; Amihood, Amir (2015): Data Quality Matters in Recommender Systems. In Hannes Werthner, Markus Zanker, Jennifer Golbeck, Giovanni Semeraro (Eds.): Proceedings of the 9th ACM Conference on Recommender Systems. Vienna, Austria. RecSys; Association for Computing Machinery; ACM Conference on Recommender Systems; ACM Recommender Systems Conference. New York, NY: ACM, pp. 257–260.

Sarwar, Badrul M.; Karypis, George; Konstan, Joseph; Riedl, John (2002): Recommender systems for large-scale e-commerce. Scalable neighborhood formation using clustering. In : Proceedings of the fifth international conference on computer and information technology, vol. 1, pp. 291–324.

Schwarz, Gideon (1978): Estimating the Dimension of a Model. In *Ann. Statist.* 6 (2), pp. 461–464. DOI: 10.1214/aos/1176344136.

Sergis, Stylianos; Sampson, Demetrios G. (2016): Learning Object Recommendations for Teachers Based On Elicited ICT Competence Profiles. In *IEEE Trans. Learning Technol.* 9 (1), pp. 67–80. DOI: 10.1109/TLT.2015.2434824.

Shani, Guy; Gunawardana, Asela (2011): Evaluating recommendation systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor (Eds.): Recommender Systems Handbook. Boston, MA: Springer US, pp. 257–297.

Shmueli; Koppius (2011): Predictive Analytics in Information Systems Research. In *MIS Quarterly* 35 (3), pp. 553–572. DOI: 10.2307/23042796.

Song, Yading; Dixon, Simon; Pearce, Marcus (2013): A survey of music recommendation systems and future perspectives. In Mitsuko Aramaki, Mathieu Barthet, Richard Kronland-Martinet, Sølvi Ystad (Eds.): From sounds to music and emotions. CMMR; International Symposium on Computer Music Modeling and Retrieval; CMMR "Music & Emotions". Berlin: Springer (Lecture Notes in Computer Science, 7900).

Symeonidis, Panagiotis (2016): Matrix and Tensor Decomposition in Recommender Systems. In Shilad Sen, Werner Geyer, Jill Freyne, Pablo Castells (Eds.): Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16. the 10th ACM Conference. Boston, Massachusetts, USA, 15.09.2016 - 19.09.2016. New York, New York, USA: ACM Press, pp. 429–430.

Tabakhi, Sina; Moradi, Parham (2015): Relevance–redundancy feature selection based on ant colony optimization. In *Pattern Recognition* 48 (9), pp. 2798–2811. DOI: 10.1016/j.patcog.2015.03.020.

Vargas-Govea, Blanca; González-Serna, Gabriel; Ponce-Medellın, Rafael (2011): Effects of relevant contextual features in the performance of a restaurant recommender system. In Bamshad Mobasher, Robin Burke, Dietmar Jannach, Gediminas Adomavicius (Eds.): Proceedings of the fifth ACM conference on Recommender systems. Proceedings of the fifth ACM conference on Recommender systems. Chicago, Illinois, USA. New York, NY: ACM, pp. 592–596.

Wand, Yair; Wang, Richard Y. (1996): Anchoring data quality dimensions in ontological foundations. In *Commun. ACM* 39 (11), pp. 86–95. DOI: 10.1145/240455.240479.

Wang, Richard Y.; Storey, Veda C.; Firth, Christopher P. (1995): A framework for analysis of data quality research. In *IEEE transactions on knowledge and data engineering* 7 (4), pp. 623–640.

Woodall, Philip; Borek, Alexander; Gao, Jing; Oberhofer, Martin; Koronios, Andy (2015): An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics. In Richard Wang (Ed.): Big data. Management and data quality ; 19th International Conference on Information Quality (ICIQ 2014), Xi'an, China, 1 - 3 August 2014. International Conference on Information Quality; ICIQ. Red Hook, NY: Curran, pp. 24–33.

Zapata, Alfredo; Menéndez, Víctor H.; Prieto, Manuel E.; Romero, Cristóbal (2015): Evaluation and selection of group recommendation strategies for collaborative searching of learning objects. In *International Journal of Human-Computer Studies* 76, pp. 22–39. DOI: 10.1016/j.ijhcs.2014.12.002.

Zhang, Zi-Ke; Zhou, Tao; Zhang, Yi-Cheng (2010): Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs. In *Physica A: Statistical Mechanics and its Applications* 389 (1), pp. 179–186. DOI: 10.1016/j.physa.2009.08.036.

4.2 Paper: Global Reconstruction of Language Models with Linguistic Rules – Explainable AI for Online Consumer Reviews

Current Status	Citation
This paper was **under review** at *30th European Conference on Information Systems* with the former title "Reconstructing the Language Model BERT with Linguistic Rules – Explainable AI for Online Consumer Reviews". After rework, this paper is **submitted** to the special issue on *Explainable and responsible artificial intelligence* of the journal *Electronic Markets* in May 2022.	Binder, M., B. Heinrich, M. Hopf and A. Schiller (2022a). "Global Reconstruction of Language Models with Linguistic Rules – Explainable AI for Online Consumer Reviews" Working Paper, University of Regensburg.

Global Reconstruction of Language Models with Linguistic Rules – Explainable AI for Online Consumer Reviews

Markus Binder, University of Regensburg, Regensburg, Germany, markus1.binder@ur.de
Bernd Heinrich, University of Regensburg, Regensburg, Germany, bernd.heinrich@ur.de
Marcus Hopf, University of Regensburg, Regensburg, Germany, marcus.hopf@ur.de
Alexander Schiller, University of Regensburg, Regensburg, Germany, alexander.schiller@ur.de

Abstract

Analyzing textual data by means of AI models has been recognized as highly relevant in information systems research and practice, since a vast amount of data on e-commerce platforms, web portals or social media is given in textual form. Here, language models such as Google's BERT, which are deep learning AI models, constitute a major breakthrough and achieve leading-edge results in many applications of text analytics such as sentiment analysis in online consumer reviews. However, these language models are "black boxes": It is unclear how they arrive at their predictions. Yet, applications of language models, for instance, in eCommerce require checks and justifications by means of global reconstruction of their predictions. To this end, we propose a novel post-hoc XAI approach by means of linguistic rules based on NLP building blocks (e.g., part-of-speech) and analyze it on different datasets and NLP tasks. Since our approach allows for different setups, we further analyze the trade-off between comprehensibility and fidelity of global reconstructions of language model predictions. We find that our approach is indeed suited for global reconstructions of BERT's predictions in online consumer reviews and in particular allows for balanced setups with respect to the trade-off between comprehensibility and fidelity. Thus, our approach paves the way for a thorough understanding of language model predictions. In practice, our approach can assist businesses in their decision-making and support compliance with regulatory requirements. Thereby, it can improve acceptance of language models and thus support their adoption.

Introduction

Huge amounts of unstructured textual data are generated across various channels of information systems (IS) such as e-commerce platforms, review portals or social media every second (Potnis, 2018). Consequently, the need for techniques that automatically analyze textual data is increasing: Until 2028, the revenues from the natural language processing (NLP) market worldwide are expected to increase at a compound annual growth rate of almost 30% to over 100 billion USD, with text analytics expected to have the highest growth (Fortune Business Insights, 2021). To

that end, the rising needs for enhanced user support as well as for understanding users' requirements are key drivers (Fortune Business Insights, 2021). As text analytics facilitate diverse applications such as sentiment analysis or text summarization (Young, Hazarika, Poria, & Cambria, 2018), various organizations in different business areas benefit from techniques of text analytics (Coheur, 2020; Zhang, Yang, Lin, & others, 2020). For instance, product or service providers can use such techniques to analyze consumer sentiments in large amounts of online consumer reviews. Using this consumer feedback enables organizations to effectively improve their products and services (Chatterjee, 2019; Heinrich, Hopf, Lohninger, Schiller, & Szubartowicz, 2019).

The state-of-the-art techniques of text analytics are language models, such as the popular deep learning AI model 'Bidirectional Encoder Representations from Transformers' (BERT) (Devlin, Chang, Lee, & Toutanova, 2019), as they have achieved leading-edge results in many tasks of text analytics (Wang et al., 2018). Language models enable a contextualized representation of text by assessing the conditional probability of each word given the contextual words surrounding it (Peters, Neumann, Iyyer, et al., 2018). Since the language model BERT is already incorporated in a plethora of business IS applications, we focus on BERT as leading exponent of language models in this paper. Amongst others, popular application scenarios of BERT in electronic markets are eCommerce, chatbots, finance or online recruiting (Coheur, 2020; Dastin, 2018; Luo, Lau, Li, & Si, 2022; Repke & Krestel, 2021; Shrestha, Krishna, & Krogh, 2021; S. Xu, Barbosa, & Hong, 2020; Yang, Uy, & Huang, 2020; Zhang et al., 2020). However, similar to most other state-of-the-art deep learning models, BERT is a "black box". That is, over 100 million learned parameters (Devlin et al., 2019) and various hidden layers contribute to BERT's immense complexity, making it hardly (if at all) possible to comprehend why and how BERT arrives at its predictions (Kovaleva, Romanov, Rogers, & Rumshisky, 2019). To address the black-box nature of AI models, a vastly increasing focus on explainable AI (XAI) in IS research and practice has emerged (Adadi & Berrada, 2018; Förster, Hühn, Klier, & Kluge, 2021; Förster, Klier, Kluge, & Sigler, 2020b). Literature agrees that the need for reconstructions and justifications is urgent and a "huge open scientific challenge" (Guidotti et al., 2018). It is even expected that "algorithmic auditing and 'data protection by design' practices will likely become the new gold standard for enterprises deploying machine learning systems" (Casey, Farhangi, & Vogl, 2019). In particular, algorithmic auditing is highly relevant for domain experts, managers and data scientists that utilize the language models' predictions for business-critical decisions or implementations and need to justify their actions. This is especially the case for application scenarios (AS) in electronic markets, as exemplarily outlined in the following and captured later on:

- eCommerce (AS1): In eCommerce, BERT is used to conduct sentiment analyses of online consumer reviews on online platforms such as Airbnb, Yelp or TripAdvisor for product development, services offerings and forecasting future demand (Heidari & Rafatirad, 2020; Shrestha et al., 2021; S. Xu et al., 2020). Since these

analyses and decisions based on BERT's predictions have large impacts, they require additional validation checks and justifications, far beyond measuring only the prediction accuracy of BERT. For instance, it needs to be ensured that specific groups of customers are not discriminated against by assigning a negative sentiment to certain countries, ethnicities or genders. Furthermore, regulations such as the General Data Protection Regulation (GDPR) in the European Union impose an extensive 'right to explanation' for automated data processing systems in general and thereby lay the foundation to enforce algorithmic auditing in companies (Casey et al., 2019).

- Chatbots (AS2): In applications in consumer services (Luo et al., 2022), BERT-based chatbots conduct direct consumer interaction and embody the company's voice. Thereby, reconstructions and justifications regarding the underlying BERT model are mandatory to prevent unhelpful, rude or misleading dialogues and thus, to support consumer satisfaction.
- Financial applications (AS3): FinBERT, which is a BERT model specifically adapted for financial text processing (Yang et al., 2020), enables the extraction of financial entities, sentiments and their relations from texts such as social media posts (e.g., tweets from CEOs or other experts) or contract documents. The extracted information is used for key tasks in finance such as accounting, auditing, compliance and risk assessment. Furthermore, language models enable to automatically process millions of documents as contained in data leaks such as the Panama Papers (O'Donovan, Wagner, & Zeume, 2019) for tax fraud detection or the analysis of large amounts of textual data regarding potential credit debtors for credit risk assessment. In particular, if legal actions are initiated based on predictions from language models (e.g., tax prosecution based on data leaks) validation checks are mandatory. Therefore, justifications are indispensable for the usage of language models for financial applications.
- Online recruiting (AS4): Supporting text analytics of application documents (Schiller, 2019), language models such as BERT enable pre-processing and pre-filtering of applications and candidates on online job platforms. Here, auditing and validation are required as such automated recruitment may lead to discrimination (e.g., by gender or origin; Dastin, 2018). Reconstructions of models help to avoid such discriminations.

These application scenarios show that it is crucial to reconstruct BERT's predictions to be able to justify the decisions based thereon. Here, the reconstructions and explanations in these scenarios are required on a global level as in all those application scenarios the predictions of language models are used in ongoing operations on a daily basis. This means that multiple decisions are made based on these predictions day-by-day for newly-generated and hitherto unknown textual data (e.g., chatbots or review summarizations are applied in real-time on consumer texts). Therefore, it is not feasible to use local approaches for reconstruction, as this would require huge efforts for manual

checks of each local reconstruction and could practically only be done a-posteriori if at all. Moreover, for instance, BERT's aspect term detections for new product or service names cannot be explained by existing local reconstructions. Therefore, global approaches are essential for reconstructions in many applications as they allow to justify predictions of a language model in advance and with reduced efforts by focusing on a global reconstruction model.

A promising way to obtain such a reconstruction and thus justify BERT's predictions is to conduct a rule-based XAI approach. On the one hand, rules are highly concrete, which also has been emphasized by Förster, Klier, Kluge, and Sigler (2020a) as decisive XAI characteristic. Indeed, studies have shown that users "prefer, trust and understand rules better than alternatives" (Ribeiro, Singh, & Guestrin, 2018; cf. also Arrieta et al., 2020). On the other hand, rule-based approaches preserve the AI model itself and thus, its high performance, while offering post-hoc reconstructions for explanations (Adadi & Berrada, 2018). Here, local rule-based approaches focus on explaining each prediction for a specific input separately, for instance, by using specific words to predict the sentiment term in a single sentence of an online consumer review. In contrast, global approaches aim at reconstructing the model's predictions as a whole (Danilevsky et al., 2020). A global approach requires a smaller rule set for reconstructing multiple predictions of a language model compared to local approaches that establish a separate and highly specific rule for each individual prediction and therefore are not really generalizable (Danilevsky et al., 2020).

To enable such a global approach, our idea is to build rules based on linguistic information (so-called linguistic rules) which generalize specific words and sentences and can be modeled by NLP building blocks such as part-of-speech tags or dependency relations (Qi, Zhang, Zhang, Bolton, & Manning, 2020). Using NLP building blocks instead of single words as rule arguments is promising for global reconstruction, as they allow for rule arguments and rules analyzing (much) more than, for instance, one single sentence in an online consumer review. Moreover, NLP relation building blocks allow to account for the contextual information in a sentence (i.e., relations between words), which is crucial for the reconstruction of language models, since language models also use contextual information. Thus, we focus on the following main research question:

> **RQ1:** How can language model predictions be globally reconstructed by means of an approach based on linguistic rules?

Analogous to local reconstructions, a global reconstruction has to be analyzed regarding its fidelity (Danilevsky et al., 2020; Gilpin et al., 2018) and comprehensibility (Guidotti et al., 2018). In case of rule-based approaches, the comprehensibility of the rule set depends on the complexity (with respect to the length of the rules; cf. Guidotti et al., 2018) and the generalizability (words vs. NLP building blocks as discussed above) of the rules. Thereby, the

comprehensibility of the rule set in our approach is adaptable (e.g., by varying rule length), which is in general outlined as an important requirement of an XAI approach (Gilpin et al., 2018). This enables to establish a balanced setup between these two objectives in a reconstruction, which further supports adoption in IS. Thus, the second research question is as follows:

> **RQ2:** How can the trade-off between fidelity and comprehensibility of global reconstructions of language model predictions by linguistic rules be analyzed?

Hence, our contribution is twofold: (1) We are the first to propose a global XAI approach for reconstructing predictions of language models by linguistic rules. In particular, (2) this paper is thus the first to analyze the trade-off between fidelity and comprehensibility (i.e., complexity and generalizability) in this setting.

For our analysis, we focus on the highly relevant tasks of aspect term detection and sentiment term detection in online consumer reviews. To that end, we use two recognized online consumer review datasets from the domains of laptops and restaurants to account for different types of goods (i.e., laptops as *search* goods and restaurants as *experience* goods). We find that our linguistic rules are indeed suited for a global reconstruction of BERT's predictions in online consumer reviews and in particular allow for balanced setups with respect to the trade-off between comprehensibility and fidelity of the reconstruction.

The remainder of this paper is structured as follows. The next section presents the background of our research. Subsequently, we discuss how to globally reconstruct language models such as BERT with linguistic rules. Thereafter, we analyze different global reconstructions of BERT, discuss their results and outline implications for research and practice. Finally, we summarize the paper and provide an outlook on future research directions.

Background

In this section, we first outline which different types of XAI approaches exist in the context of language models. Second, several NLP building blocks recognized by literature are introduced forming the basis for our approach (contribution (1)). The section concludes with a discussion of related work yielding the addressed research gap.

Types of XAI approaches in the context of language models

To clarify the notion of XAI (i.e., what explainable AI really means), a characterization in opaque systems, interpretable systems and comprehensible systems has been proposed (Doran, Schulz, & Besold, 2017). Here, *opaque systems* offer no insights into the system's reasoning on how inputs are mapped to the corresponding outputs. In that line, modern language models such as BERT are opaque systems, as it is not possible to comprehend

its mappings, for instance, comprising over 100 million learned parameter values in the case of BERT. Based on that, there are two separate notions of addressing this problem. First, *interpretable systems* allow to understand how inputs are mapped to outputs by subdividing the mapping. This is not feasible for language models such as BERT due to its large amount of parameters and layers, which results in highly complex concatenated functions (Devlin et al., 2019). Second, *comprehensible systems* allow to relate *properties* of the inputs, for instance, single terms of an input sentence, to their output such as a classification of sentiment terms (Doran et al., 2017). While research in both areas is important, it has to be pointed out that the resulting XAI approaches are not "actually" explanation systems (Doran et al., 2017). For instance, rule-based approaches mostly give insights on how, but not why specific predictions are made (Doran et al., 2017). That is, causality cannot be directly established. To account for these different notions, we deliberately refer to "reconstructing" BERT rather than "explaining" in this paper.

Related to the two notions of interpretable and comprehensible systems, there are, in general, two main approaches in XAI (Adadi & Berrada, 2018): On the one hand, *intrinsic XAI approaches* 'force' the AI model (during training) to produce interpretable mappings from input to output (Adadi & Berrada, 2018). The drawback of these intrinsic approaches is that they are limited in the type of interpretations they can provide, as they need to restrict the model to obtain interpretable mappings, thus usually worsening the model's performance (Adadi & Berrada, 2018). Due to its complexity, BERT would have to be extremely simplified to enable interpretable mappings. On the other hand, *post-hoc XAI approaches* aim to comprehensibly reconstruct the mappings from input to output of an AI model. These approaches do not require to restrict the model during training (Adadi & Berrada, 2018). Here, a popular method is rule extraction, since rules can potentially exhibit a high degree of comprehensibility (Ribeiro et al., 2018). In general, there are two categories of rule extraction techniques (Adadi & Berrada, 2018):
1) *Decompositional rule extraction* aims at extracting rules at selected, often single nodes within a neural network. To comprehend the predictions of a language model, it is then necessary to concatenate multiple extracted rules for various hidden layers. Thus, the drawback of this technique is that concatenations of rules are highly complex for deep neural networks such as BERT (Augasta & Kathirvalavakumar, 2012). Since the resulting rules would again be difficult to comprehend, decompositional rule extraction is not feasible for comprehensibly reconstructing language models. 2) In contrast, *pedagogical rule extraction* aims at extracting rules considering only the inputs and outputs. In particular, rules are extracted based on properties of the inputs and the corresponding outputs to reconstruct the mappings of the AI model. Thus, this approach can contribute to a comprehensible reconstruction even for language models such as BERT, since the extracted rules do not have to be concatenated through the various hidden layers.

Additionally, a further important differentiation within post-hoc XAI research is between *global* and *local approaches* (Danilevsky et al., 2020). Here, global approaches aim at reconstructing the predictions of an AI model

by means of one single global model (Danilevsky et al., 2020). In contrast, local approaches create separate, highly specific reconstruction models for each prediction (e.g., in a single sentence of an online consumer review). To enable local reconstructions for IS text analytics applications, rules solely based on specific words are used by extant literature (e.g., Ribeiro et al., 2018). However, such rules lack the ability to generalize. In contrast, linguistic rules based on NLP building blocks are more promising for the global reconstruction of language models. Indeed, rule arguments with NLP building blocks generalize much better than rule arguments with specific words, and NLP relation building blocks enable to incorporate contextual information, which is a main component of language models.

Both objectives *fidelity* and *comprehensibility* are crucial for global post-hoc XAI approaches (Arrieta et al., 2020; Guidotti et al., 2018; Szczepański, Pawlicki, Kozik, & Choraś, 2021). Indeed, on the one hand, a global reconstruction needs to match the predictions of an AI model to avoid false conclusions, which is measured by fidelity (Gilpin et al., 2018). On the other hand, comprehensibility (i.e., complexity and generalizability; commonly measured in terms of model size) enables the use of the reconstruction (Guidotti et al., 2018). Thus, we analyze the reconstruction of BERT regarding its fidelity and its comprehensibility and strive to enable different setups between the two objectives.

NLP building blocks

To enable a reconstruction using linguistic rules, our idea is to use different semantical and syntactical NLP building blocks (cf. Introduction). Thus, we briefly outline NLP building blocks that are widely recognized in the literature (Fellbaum, 2013; Kamps, Marx, Mokken, & Rijke, 2004; Tenney, Xia, et al., 2019) and that constitute a basis for our reconstruction. Table 1 summarizes these different building blocks. Thereby, the column 'type' characterizes a

Building block	**Type**	**Linguistic information**	**Example labels for the sentence "The waiter of The Burger House was nice, he smiled at us."**
Part-of-speech tags (POS)	Tags	Syntactic	POS-label ("*waiter*") = NN (Noun)
Synsets (SYN)	Tags	Semantic	SYN-label ("*nice*") = nice.a.01 (Synset description: "pleasant or pleasing or agreeable in nature or appearance")
Dependencies (DEP)	Relations	Syntactic	DEP-label ("*waiter*", "*nice*") = amod (adjectival modifier)
Semantic role labeling (SRL)	Relations	Semantic	SRL-label ("*he*", "*smiled*") = agent-predicate-relation
Coreferences (COREF)	Relations	Semantic	COREF-label ("*waiter*", "*he*") = True (referring to the same entity)
Proximity (PROX)	Relations	Syntactic	PROX-label ("*waiter*", "*nice*") = 6

Table 1. Overview of NLP building blocks.

building block as *tag* or *relation* (as described in the following). In addition, the column 'linguistic information' shows whether a building block provides semantic or syntactic information. For each building block, an example is given in the last column.

A *tag building block* provides tag labels for selected tokens (e.g., words) of a sentence. Tag labels describe a certain syntactic or semantic information of tokens in consideration of the whole sentence. Part-of-speech (POS) tags provide information on the *syntactic* structure of a sentence. Thereby, the POS tag, such as noun, adjective or verb, is assigned to a single token. The building block synsets (SYN) considers the *semantic* information of tokens. In particular, SYN labels (e.g., derived from the lexical database WordNet) indicate words which share the same or a similar meaning (Fellbaum, 2013) taking into account its word context in a sentence.

A *relation building block* provides a label for a pair of tokens in a sentence describing a certain syntactic or semantic relation between these tokens. These relation building blocks enable to account for the contextual information in a sentence (i.e., the relation between tokens in a sentence), which is crucial for a reconstruction of BERT as BERT also considers contextual information. A basic *syntactic* information is the distance between two tokens, which is covered by the proximity (PROX) building block. For instance, if two tokens are next to each other in a sentence, their distance is 1. Dependencies (DEP) also link two tokens based on their *syntactical* relationship (e.g., adjectival modifier or nominal subject) (Manning et al., 2014). *Semantic* information is provided by the building blocks semantic role labeling (SRL) and coreference (COREF). SRL relations identify combinations of predicates and semantic arguments in a sentence (Tenney, Xia, et al., 2019). COREF links two tokens referring to the same entity (Tenney, Xia, et al., 2019). Consequently, information referring to one part of the relation can be traced back to the other part.

Related Work

Our goal is to reconstruct the language model BERT by means of linguistic (pedagogical) rules composed of NLP building blocks. Hence, XAI approaches analyzing language models regarding NLP building blocks (category A) as well as XAI approaches analyzing pedagogical rules for reconstructing language models (category B) constitute the related work. In contrast, general rule-based XAI approaches (cf. Adadi & Berrada, 2018) and XAI approaches (Ramon, Martens, Evgeniou, & Praet, 2020; Sushil, Šuster, & Daelemans, 2018) relying on a simple 'bag-of-words' analysis – both without any focus on language models – are not in the scope for our research.

Ad category A): Several existing works analyze language models by using their (contextualized) word embeddings or internal states as input to *predict* NLP building blocks (Coenen et al., 2019; Hewitt & Manning, 2019; Jumelet & Hupkes, 2018; Kim, Patel, Poliak, Wang, & others, 2019; Peters, Neumann, Zettlemoyer, & Yih, 2018; Tenney,

Das, & Pavlick, 2019; Tenney, Xia, et al., 2019; van Aken, Winter, Löser, & Gers, 2019). Then, the quality of these predictions is used as an indication whether a certain NLP building block is encoded in particular word embeddings (i.e., vector representations) or specific layers of the language models. That is, instead of reconstructing predictions of language models for NLP tasks in IS (e.g., sentiment term detection), an analysis of the general word embeddings themselves is aimed for in these works. For instance, different NLP building blocks have been predicted by word embeddings of the language models ELMo (Peters, Neumann, Zettlemoyer, & Yih, 2018) and BERT (Tenney, Das, & Pavlick, 2019; Tenney, Xia, et al., 2019). However, the aim of our research is a different one. As discussed in the Introduction, our focus is to better comprehend BERT's predictions on NLP tasks in IS, for instance, to be able to justify decisions made based on its results. To enable that, it is necessary to reconstruct the predictions of BERT for relevant NLP tasks (such as the extracted sentiment terms in online consumer reviews), since these predictions and not particular word embeddings in form of vector representations are the foundation for further decisions. In that line, none of the approaches in this category considers pedagogical rules to enable a reconstruction of predictions of a language model for NLP tasks in IS.

Ad category B): There also exist recent, interesting works that analyze language models by means of pedagogical rules in a local manner (i.e., for single predictions). In Ribeiro et al. (2018), individual predictions of simple recurrent neural network-based language models are reconstructed by separate if-then rules. Building on this work, BERT's predictions in an application of fake news detection on social media are analyzed in Szczepański et al. (2021). Both works hardly incorporate contextual information for reconstructions. That is, only information of the previous token is considered to obtain local reconstruction rules. Thus, both works consider only short rules of low complexity. In addition, rules based on individual tokens (e.g., specific words) are used. Hence, both works do not discuss the composition of tag and relation building blocks when extracting rules for reconstruction and as a result, the proposed rules exhibit only low generalizability. In particular, relation building blocks such as DEP or COREF, which enable rules to comprise vital contextual information, are not considered.

A detailed categorization of existing approaches for reconstruction of language models based on pedagogical rules (category B) is further illustrated in Table 2. Overall, existing rule-based reconstructions give (highly) limited

	Local / Low Generalizability	**Global / High Generalizability**
Low Complexity	Ribeiro et al. (2018); Szczepański et al. (2021)	-
High Complexity	-	-

Table 2. Categorization of existing approaches for the reconstruction of language models by means of pedagogical rules.

insights on how predictions are made by language models in general, impeding their adoption for IS tasks.

Overall, while the approaches in category A) give interesting indications on how NLP building blocks may be encoded in contextualized word embeddings, they do not enable to reconstruct the predictions of language models in NLP tasks in IS. In contrast, the approaches in category B) indeed analyze pedagogical rules for reconstructing specific predictions, but only enable local reconstructions and do not incorporate different NLP building blocks comprising contextual linguistic information.

Summing up, there are very interesting contributions in the field of XAI regarding language models. However, literature lacks approaches for global reconstructions of the predictions of language models for NLP tasks in IS (e.g., sentiment term detection in online consumer reviews) based on pedagogical rules (cf. Table 2). To address this research gap, this paper proposes, to the best of our knowledge, the first global XAI approach for reconstructing predictions of language models by linguistic (pedagogical) rules. In particular, this paper is thus the first to enable an analysis of the trade-off between fidelity and comprehensibility (i.e., complexity and generalizability) in this setting.

Global Reconstruction of BERT with Linguistic Rules

In this section, we introduce our approach by postulating the formal structure of linguistic rules for the global reconstruction of BERT predictions and then outline appropriate measures to analyze this reconstruction.

Formal structure of linguistic rules for reconstructing BERT

We begin by deriving the formal structure of linguistic rules. Thereby, for illustration purpose, the language model BERT is applied for the token classification tasks aspect term detection and sentiment term detection that are frequently used in online consumer reviews (Dai & Song, 2019; Sun, Huang, & Qiu, 2019; H. Xu, Liu, Shu, & Yu, 2019). More precisely, each sentence in a document comprises a string value and can be split up by tokenization into disjunct substrings (so-called tokens), which have a linguistic meaning, such as (sub)words or punctuation marks. The precise tokenization of sentences depends on specific tokenization policies. For this work, we used w. l. o. g the widely-applied tokenization of the python package NLTK (cf. https://www.nltk.org). The goal of the token classification tasks performed by BERT is to assign class labels to such tokens. For example, the token ‘fish’ in the sentence ‘The fish was good!’ is assigned with the class label ASP indicating an aspect term. The following postulates P1)-P3) provide the foundation for linguistic rules based on NLP building blocks, which enable a global reconstruction of BERT's predictions (i.e., the predicted class labels for the tokens of a sentence).

P1) “LABEL ASSIGNMENTS”: In this work, we assign labels only to single tokens or token pairs. Hence, we do not consider label assignments for whole sentences, documents nor for single character values.

P1.1) "TAG LABEL ASSIGNMENTS": A tag building block $tbb \in TBB$ (where TBB is the set of tag building blocks) assigns at most one *tag label* $l_{tbb}(t_i) \in L_{tbb}$ to a token t_i (L_{tbb} is the set of all labels from tbb). For example, the tag building block POS with L_{POS} ={"NN", "VB", "JJ", …} assigns the label $l_{POS}(t_2)$ ="NN" to the token t_2 ='fish'.

P1.2) "RELATION LABEL ASSIGNMENTS": A relation building block $rbb \in RBB$ (where RBB is the set of relation building blocks) assigns at most one *relation label* $l_{rbb}(t_i, t_j) \in L_{rbb}$ to a token pair (t_i, t_j) (L_{rbb} is the set of all labels from rbb). For example, the relation building block DEP with L_{DEP} ={"amod", "nsubj",…} assigns the label $l_{DEP}(t_2, t_4)$ ="nsubj" to the token pair (t_2, t_4)=('fish', 'good').

P1.3) "CLASS LABEL ASSIGNMENTS": BERT assigns a *class label* $l_\tau(t_i) \in L_\tau$ to each token t_i (L_τ is the set of all class labels in a token classification task). For instance, in the aspect term detection task with class labels $L_{asp} = \{ASP, \overline{ASP}\}$, the token t_2 ='fish' is assigned with the class label ASP by BERT indicating that 'fish' is an aspect term.

P2) "FEASIBLE ARGUMENTS FOR RULES": In this work, *feasible arguments* in the antecedent and consequents of a rule only reference to labels for tokens or token pairs as postulated in P1).

P2.1) "FEASIBLE ARGUMENTS IN RULE ANTECEDENTS": In this work, a feasible argument in the rule antecedent only contains conditions regarding tag labels of tokens (cf. P1.1)) and relation labels of token pairs (cf. P1.2)).

P2.2) "FEASIBLE ARGUMENTS IN RULE CONSEQUENTS": In this work, a feasible argument in the rule consequent only contains class label assignments of tokens (cf. P1.3). Considering the classification task of sentiment term detection, the argument $l_{SENT}(t_4) \rightarrow SENT$ assigns the label $l = SENT$ to the token t_4 ='good', indicating that 'good' is labelled as a sentiment term by BERT in the sentence 'The fish was good!'.

P3) "CONFLICTING CLASSIFICATION RESULTS OF MULTIPLE RULES": Multiple rules $R_1, \dots, R_{n_R}$ ($n_R \in \mathbb{N}$) may result in conflicting classification results $l_1(t_i), \dots, l_{n_R}(t_i) \in L_\tau$ for the same token t_i. To resolve such conflicting classification results for a token t_i, it is sensible to assign the class of the rule with the highest precision (cf. next section).

Given the postulates P1)-P3), the structure of linguistic rules can be defined. A linguistic rule R is an "if-then-else" rule in the form of **IF** antecedent **THEN** "then"-consequent (**ELSE** "else"-consequent). Here, the antecedent is an arbitrary combination of feasible arguments as postulated in P2.1) by means of logical operators such as AND (i.e., "∧"), OR (i.e., "∨") and NOT (i.e., "¬"). Further, each "then"-consequent and each "else"-consequent consists of one feasible argument as postulated in P2.2). Thus, a rule R outputs the class assignments of the "then"-consequent

in case that the antecedent is TRUE (otherwise and if an "else"-consequent is contained in the rule, it outputs the class assignments of the "else"-consequent). Moreover, rules can be characterized by their length, which is given by the number of tokens that are connected by a relation building block in the antecedent of a rule. A brief example of a simple rule of length two is given by:

IF $\left([POS(t_i) == NN] \vee \neg\left[POS(t_j) == VB\right]\right) \wedge \left[DEP(t_i, t_j) == nsubj\right]$

THEN $l_{ASP}(t_i) \rightarrow ASP$

This rule can be applied to the tokenized sentence ('The', 'fish', 'was', 'good', '!') from above. For this sentence, the antecedent of the rule is only TRUE if $t_i = t_2 =$'fish' and $t_j = t_4 =$'good'. For any other selection of t_i and t_j, the antecedent is FALSE since only the token pair ('fish', 'good') has the relation "nsubj" in this sentence. Hence, this linguistic rule correctly detects the aspect term 'fish'. Rules of the outlined formal structure based on the postulates P1)-P3) constitute the foundation for reconstructing BERT in this work.

Assessing fidelity and comprehensibility of global reconstructions

To globally reconstruct BERT, all predictions of BERT for a token classification task have to be considered. Here, fidelity and comprehensibility are the most relevant measures (cf. Section "Types of XAI approaches in the context of language models") and assessing both measures is required to analyze the trade-off between fidelity and comprehensibility. Since we focus on global reconstructions of language models, we outline in detail how both measures can be assessed for global reconstructions in the following.

To measure fidelity, we consider the predictions of BERT for each class label. More precisely, the set of token ids (i.e., the positions of tokens in the text corpus) predicted by BERT as class $C \in L_\tau$ is given by $I_{C,BERT} = \{i \in I | l_{BERT}(t_i) = C\}$. These token ids are used as the basis for extracting the linguistic rules on training data $I_{train,C,BERT}$ and validation data $I_{validation,C,BERT}$ as well as for assessing their fidelity of globally reconstructing BERT on test data $I_{test,C,BERT}$. Once a set Σ of linguistic rules is extracted, the F1 score is appropriate to assess the fidelity of the rule set (Sushil et al., 2018) as – in contrast to the accuracy measure – it accounts for imbalanced class distributions. The F1 score (i.e., based on precision and recall) of the rule set Σ for reconstructing BERT's predictions $I_{C,BERT}$ is given by:

$$Pr_C(\Sigma) = \frac{\left|I_{test,C,BERT} \cap I_{test,C,\Sigma}\right|}{\left|I_{test,C,\Sigma}\right|} \quad (1)$$

$$Rec_C(\Sigma) = \frac{\left|I_{test,C,BERT} \cap I_{test,C,\Sigma}\right|}{\left|I_{test,C,BERT}\right|} \quad (2)$$

$$F1_C(\Sigma) = \frac{2 * Pr_C(\Sigma) * Rec_C(\Sigma)}{Pr_C(\Sigma) + Rec_C(\Sigma)} \quad (3)$$

Here, $I_{test,C,\Sigma} = \{i \in I_{test} | l_{rule}(\Sigma, t_i) = C\}$ is the set of token ids from the test data that are assigned with class C by the rule set Σ. In case of multiclass classification the fidelity is then assessed by the average F1 score per class label C, denoted as $\overline{F_1}(\Sigma)$ (i.e. by the macro-averaged F1 score (Sushil et al., 2018)). In contrast to the regular formulas for classifier evaluation, which aim to evaluate the predictions of a classifier regarding the true class labels, the formulas (1)-(3) enable to evaluate the linguistic rules regarding the predicted class labels by BERT and hence, to assess the fidelity of reconstructing BERT by certain sets of linguistic rules Σ.

In contrast to the comprehensibility of local reconstructions (e.g., complexity and generalizability of single rules), literature (Guidotti et al., 2018) suggests to assess the comprehensibility of a global reconstruction by its model size. Since our model is a set of rules Σ, the comprehensibility of this global reconstruction can be measured by the number of rules $NR(\Sigma)$ in the rule set and the number of unique argument values $NUAV(\Sigma)$ in the antecedents in the rule set (Vilone & Longo, 2021), as given by:

$$NR(\Sigma) = |\Sigma| \quad (4)$$

$$NUAV(\Sigma) = |\{v \in AAV | \exists R \in \Sigma: v \in R\}| \quad (5)$$

Here, $AAV = L_{POS} \cup L_{SYN} \cup L_{DEP} \cup L_{SRL} \cup L_{COREF} \cup L_{PROX}$ is the set of all argument values of all NLP building blocks.

Analysis

In this section we analyze the reconstruction of BERT's predictions by our approach. First, we outline the selected tasks, datasets and the conducted automated extraction of linguistic rules for global reconstruction. Then, we demonstrate how our approach based on linguistic rules can reconstruct predictions of BERT. After that, we present and discuss the results as well as implications for research and practice.

Task selection, data preparation and rule extraction

For a meaningful analysis of the reconstruction of BERT's predictions, we selected the NLP tasks aspect term detection and sentiment term detection as these tasks are frequently analyzed in the IS field and constitute common applications for BERT and text analytics (Dai & Song, 2019; Sun et al., 2019; H. Xu et al., 2019), in particular in

electronic markets (Chatterjee, Goyal, Prakash, & Sharma, 2021; Steur, Fritzsche, & Seiter, 2022). Also, we chose two publicly available datasets that exhibit different characteristics – with restaurants reviews from the platform *Yelp* (Yelp Dataset Challenge; cf. https://www.yelp.com/dataset) as experience goods vs. laptop reviews from the platform *Amazon* (Ni, Li, & McAuley, 2019) as search goods – to enable broader insights independent of specific item domains. To extract linguistic rules based on the formal structure postulated in the previous section, we extended existing techniques from the literature.

Dataset characteristic	**Restaurants**	**Laptops**
# of sentences	70,000	70,000
# of tokens	1,080,347	1,286,432
# of predicted aspect tokens by BERT	92,853	120,993
# of predicted sentiment tokens by BERT	90,384	72,615
Relative frequency of predicted aspect tokens by BERT (relative to # of tokens or # of sentences)	0.086 (rel. to tokens); 1.326 (rel. to sentences)	0.094 (rel. to tokens); 1.728 (rel. to sentences)
Relative frequency of predicted sentiment tokens by BERT (relative to # of tokens or # of sentences)	0.084 (rel. to tokens); 1.291 (rel. to sentences)	0.056 (rel. to tokens); 1.037 (rel. to sentences)

Table 3. Datasets for analysis.

In detail, the goal of aspect term detection and sentiment term detection is to classify tokens in online consumer reviews that express aspects or sentiments. An aspect term (e.g., 'laptop screen') represents an item aspect for which an opinion polarity is expressed by a sentiment term (e.g., 'very good') (Sun et al., 2019). The task of token classification is to assign a class label $C \in L_\tau$ (i.e., $L_{asp} = \{ASP, \overline{ASP}\}$ and $L_{sent} = \{SENT, \overline{SENT}\}$) to tokens of a sentence. To conduct aspect term detection and sentiment term detection, we used the publicly available state-of-the-art post-trained BERT models (H. Xu et al., 2019). That is, the tokens of both datasets were assigned with the class labels of BERT's predictions. An overview of the (randomly sampled) dataset excerpts used for analysis, including the predictions of BERT regarding both tasks, is given in Table 3.

To prepare the datasets for the analysis, we randomly split the sentences of the datasets into 65% training data, 15% validation data and 20% test data. Then, the extraction of linguistic rules comprises two steps. Firstly, automated rule generation determines linguistic rules that appear at minimum five times in the training data to avoid rules that are only applicable for very few and highly specific sentences. Secondly, the rule selection assembles a subset of these linguistic rules by iteratively adding rules to a (initially empty) rule set if the F1 score of the rule set is thereby enhanced on the validation data (Liu, Gao, Liu, & Zhang, 2015). To conduct the extraction of rules, we extended existing techniques for automated rule generation (Dai & Song, 2019) and automated rule selection (Liu et al., 2015) to enable an integration and combination of different NLP building blocks. Then, we assessed the F1 score of the extracted set of linguistic rules on the test data. To assess comprehensibility of the extracted rules, we focused on rules with antecedents containing *at most* two arguments regarding tag building blocks and *at most* one argument

regarding a relation building block. Hence, the rules are of at most length two. In that line, we only used the logical operator "AND" to preserve comprehensibility (Askira-Gelman, 1998).

Demonstration of reconstructing BERT's predictions with linguistic rules

Before discussing the results based on the introduced datasets and tasks, we give a brief preliminary demonstration of how our approach based on linguistic rules can be utilized to reconstruct predictions of BERT. Thereby, we consider the following three exemplary sentences of real restaurant reviews and highlight the extracted sentiment terms of BERT by bold font: "*The Homeburger was **huge**.*", "*Moreover, John is **friendly** and **welcoming**.*", "*Overall, the BurgerBarn is **amazing**.*". A linguistic rule proposed by our approach that reconstructs these predicted sentiment terms is given by:

IF $[POS(t_i) == NNP] \wedge [POS(t_k) == JJ] \wedge [DEP(t_i, t_k) == nsubj]$

THEN $l_{SENT}(t_k) \rightarrow SENT$

This single rule detects the adjectives, which are in a nominal subject relation ("nsubj") with a proper noun ("NNP"), as sentiment terms. The application of this rule for the three sentences is given in Table 4.

Example sentence	**Application of the above linguistic rule**
"*The Homeburger was **huge**.*"	**IF** $[POS(Homeburger) == NNP] \wedge [POS(huge) == JJ] \wedge [DEP(Homeburger, huge) == nsubj]$ **THEN** $l_{SENT}(huge) \rightarrow SENT$
"*Moreover, John is **friendly** and welcoming.*"	**IF** $[POS(John) == NNP] \wedge [POS(friendly) == JJ] \wedge [DEP(John, friendly) == nsubj]$ **THEN** $l_{SENT}(friendly) \rightarrow SENT$
"*Moreover, John is friendly and **welcoming**.*"	**IF** $[POS(John) == NNP] \wedge [POS(welcoming) == JJ] \wedge [DEP(John, welcoming) == nsubj]$ **THEN** $l_{SENT}(welcoming) \rightarrow SENT$
"*Overall, the BurgerBarn is **amazing**.*"	**IF** $[POS(BurgerBarn) == NNP] \wedge [POS(amazing) == JJ] \wedge [DEP(BurgerBarn, amazing) == nsubj]$ **THEN** $l_{SENT}(amazing) \rightarrow SENT$

Table 4. Application of a linguistic rule to reconstruct BERT's predictions in exemplary sentences.

As illustrated in Table 4, the rule reconstructs the sentiment terms detected by BERT in these example sentences and constitutes a generalizing, plausible rule, which is important for online consumer reviews, as special product/service names or attributes (e.g., special dishes or waiters in restaurant reviews) are often referenced by proper nouns. Overall, this rule alone already reconstructs around 350 sentiment terms in the restaurant dataset with a precision of 89% with respect to BERT's predictions. In contrast, reconstructing these sentiment terms by means of rules with

specific tokens instead of NLP tag building blocks, a separate rule for each instantiation in Table 4 would be required for each of the sentiment terms. For instance, the rule

IF $[TOKEN(t_i) == John] \wedge [TOKEN(t_k) == friendly] \wedge [DEP(t_i, t_k) == nsubj]$

THEN $l_{SENT}(t_k) \rightarrow SENT$

is obviously highly specific and cannot reconstruct the sentiment terms 'huge, 'welcoming' or 'amazing'. Therefore, this example emphasizes that linguistic rules with NLP building blocks enable to achieve higher generalizability for a reconstruction of the predictions of language models (e.g., in online consumer reviews).

Results

In this section, we present the results of the proposed approach for the reconstruction of BERT's predictions. In particular, we analyze the fidelity and the comprehensibility to which an extracted set of rules is able to globally reconstruct BERT. To account for the objectives of high fidelity and high comprehensibility, we consider four different setups of rule complexity and rule generalizability as outlined in Table 2: To analyze rule complexity, we distinguish between "L1-rules" containing rules of length one and "L2-rules" comprising rules of length at most two (i.e., every L1-rule is also a L2-rule, but not vice versa). We point out that L2-rules contain relation labels and thus consider contextual information, while this is not possible for L1-rules. To analyze rule generalizability, we compare rules with specific tokens as arguments (low generalizability) against rules with NLP building blocks and without specific token arguments (high generalizability). Given this, the comprehensibility of the four setups is shown in the Tables 5-8 regarding both tasks on the respective datasets.

Aspect Term Detection	**Restaurants**		**Laptops**	
	Low Generalizability (i.e., rules with specific tokens)	**High Generalizability** (i.e., rules with NLP building blocks)	**Low Generalizability** (i.e., rules with specific tokens)	**High Generalizability** (i.e., rules with NLP building blocks)
Low Complexity (i.e., L1-rules)	1033 / 1033	26 / 26	1006 / 1006	32 / 32
High Complexity (i.e., L2-rules)	7277 / 2384	2280 / 242	6362 / 2221	2700 / 315

Table 5. Comprehensibility of the global reconstruction of BERT's predictions for aspect term detection. Comprehensibility is measured by: NR / $NUAV$. The coloring of the cells indicates high comprehensibility (dark green), medium comprehensibility (light green and yellow) and low comprehensibility (red). The coloring depends on $NUAV$.

Aspect Term Detection	Restaurants		Laptops	
	Low Generalizability	High Generalizability	Low Generalizability	High Generalizability
Low Complexity	72.2% (71.9%,72.6%)	53.5% (39.0%,85.2%)	80.4% (83.8%,77.4%)	53.2% (38.5%,85.8%)
High Complexity	76.4% (74.4%,78.5%)	63.7% (54.6%,76.3%)	85.3% (86.2%,84.4%)	63.8% (56.9%,72.6%)

Table 6. Fidelity of the global reconstruction of BERT's predictions for aspect term detection.
Fidelity is measured by: F1 Score (Precision, Recall). The coloring of the cells indicates high fidelity (dark green), medium fidelity (light green and yellow) and low fidelity (red).

Sentiment Term Detection	Restaurants		Laptops	
	Low Generalizability	High Generalizability	Low Generalizability	High Generalizability
Low Complexity	836 / 836	17 / 17	714 / 714	14 / 14
High Complexity	4584 / 1730	1744 / 264	4201 / 1440	1718 / 300

Table 7. Comprehensibility of the global reconstruction of BERT's predictions for sentiment term detection.
Comprehensibility is measured by: NR / $NUAV$. The coloring of the cells indicates high comprehensibility (dark green), medium comprehensibility (light green and yellow) and low comprehensibility (red). The coloring depends on $NUAV$.

Sentiment Term Detection	Restaurants		Laptops	
	Low Generalizability	High Generalizability	Low Generalizability	High Generalizability
Low Complexity	76.8% (81.6%,72.5%)	66.7% (61.5%,72.8%)	74.5% (75.9%,73.2%)	56.5% (48.7%,67.4%)
High Complexity	81.4% (82.5%,80.3%)	70.3% (68.0%,72.8%)	78.1% (78.0%,78.1%)	64.6% (62.4%,67.0%)

Table 8. Fidelity of the global reconstruction of BERT's predictions for sentiment term detection.
Fidelity is measured by: F1 Score (Precision, Recall). The coloring of the cells indicates high fidelity (dark green), medium fidelity (light green and yellow) and low fidelity (red).

Discussion of the results

We elaborate on the major findings of applying our approach to globally reconstruct language model predictions by discussing the research questions RQ1 and RQ2:

Ad RQ1: *Our approach based on linguistic rules allows for the global reconstruction of language model predictions (e.g., in online consumer reviews).*

Our analysis shows that the predictions of BERT can be globally reconstructed by our approach with a fidelity of 76%-85% on the considered tasks for online consumer reviews (cf. Tables 6 and 8). In more detail, the recall of the global reconstructions (i.e., how many classified tokens of BERT could be reconstructed) based on L2-rules with

tokens ranges between 78%-84%, while the precision of these global reconstructions ranges between 74%-86%. This shows that incorporating relation building blocks capturing contextual information in the L2-rules is indeed helpful to globally reconstruct BERT's predictions with higher fidelity. In addition, the rule sets with NLP building blocks and without token arguments yield higher comprehensibility, which is indicated by a low numbers of unique argument values (below 320) compared to over 70,000 classified tokens by BERT. Here, it could be substantiated that the reconstruction of BERT's predictions is constituted by transparent rules. For instance, the rule "**IF** a term is a synset of 'good' and an adjectival modifier (DEP-relation 'amod') of a noun (POS-tag 'NN'), **THEN** that token is labelled as a sentiment term by BERT." achieved 100% precision and over 500 reconstructed terms in the restaurant dataset. In particular, no discriminating factors such as specific synsets regarding gender, origin or neglected negative sentiments for specific products/ services were detected. In combination with the decent fidelities of these comprehensible reconstructions, this yields that the reconstruction provided by our approach can be used for algorithmic auditing including validation checks in application scenarios such as the four discussed in the introduction (AS1-4). Overall – in contrast to related work – the linguistic rules in our approach enable a global reconstruction of BERT's predictions by means of NLP building blocks.

Ad RQ2: *Our approach enables to establish a balanced setup between fidelity and comprehensibility*.
As the proposed linguistic rules allow to vary their rule complexity (e.g., L2-rules vs. L1-rules) and their rule generalizability (e.g., rules with NLP tag building blocks vs. specific tokens), it is possible to create setups for global reconstructions with different comprehensibility (cf. Tables 5 and 7). Our analysis of these setups shows that higher fidelity is achieved by reducing comprehensibility and vice versa. This yields that fidelity and comprehensibility are two conflicting objectives, which has also been indicated in general in XAI literature (Arrieta et al., 2020; Gilpin et al., 2018). Indeed, the reconstruction by means of linguistic rules can either have a higher fidelity or a higher comprehensibility, while both objectives cannot be achieved simultaneously. In particular, our results show that L1-rules with NLP BBs and without token arguments, which have low complexity and high generalizability, yield the global reconstruction with the highest comprehensibility (i.e., NR and NUAV are below 35; cf. Tables 5 and 7) in comparison. These rule sets achieve fidelities between 53% to 67% (cf. Tables 6 and 8). This means that BERT's predictions on the tasks of aspect term detection and sentiment term detection can already be partly reconstructed in a very comprehensible manner with a set of rules of only one tag building block as argument. Conversely, when utilizing specific tokens instead of NLP building blocks as arguments in L1-rules, a higher fidelity of 72% to 80% is achieved (cf. rules with low generalizability in Tables 6 and 8). However, such rules (e.g., the rule "flavorful is a sentiment term") are highly specific and have low generalizability, which results in a rule set with over one thousand rules and unique argument values in the antecedents (cf. Tables 5 and 7).

Furthermore, Table 6 and Table 8 indicate that the fidelity increases when the rules become more complex, but this is accompanied by a decreasing comprehensibility as indicated in Tables 5 and 7. Here, L2-rules with NLP building blocks and without token arguments achieve fidelities between 64% to 70% (cf. Tables 6 and 8) with at most 315 unique argument values. Contrarily, L2-rules with tokens achieve the highest fidelities with values from 76% up to 85% (cf. Tables 6 and 8), but they exhibit the lowest generalizability and thus, global reconstructions with low comprehensibility which is indicated by multiple thousands of rules and between 1,400 and 2,400 unique argument values (cf. Tables 5 and 7) in the rule sets. These different setups show that either higher fidelity or higher comprehensibility can be achieved by reconstructing BERT's predictions with linguistic rules. However, if both objectives are crucial and focused equally, the best setup may be L2-rules with NLP building blocks (and without token arguments), which exhibit decent fidelity and comprehensibility at the same time. The advantage of these L2-rules compared to L1-rules with tokens is the much lower number of unique argument values, which is based on the higher generalizability of NLP building blocks compared to specific tokens, and in particular, the use of contextual information in form of relation building blocks. Overall, our linguistic rules enable to establish different relevant setups with respect to fidelity and comprehensibility depending on the requirements for an XAI approach in practice.

Implications for research and practice

Our work contributes to the comprehensibility of opaque AI models in text analytics, as it allows for comprehensible global reconstructions of language models. Therefore, our work is not only valuable for multiple different research strands, but it is also highly relevant for applications and supports the adoption of language models, as outlined in the following.

Implications for research

1) *Linguistic rules enable global reconstructions of high fidelity for language model predictions in text analytics.* Existing literature on XAI (e.g., Arrieta et al., 2020) discusses that rule-based XAI models can exhibit high comprehensibility but tend to lack high fidelity for reconstructions of complex AI models. Our findings extend this existing body of knowledge, as our analysis shows that our approach based on linguistic rules enables reconstructions with high fidelity as well as reconstructions with high comprehensibility for language model predictions. In particular, linguistic rules can achieve high fidelity by means of the contained relation building blocks capturing contextual information which is relevant for many text analytics tasks (Devlin et al., 2019; Geng, Zhang, & Han, 2021; Peters, Neumann, Iyyer, et al., 2018). However, we also find that high fidelity and high comprehensibility of rule-based reconstruction of language models cannot be reached by the same reconstruction, as

fidelity increases by reducing comprehensibility. Here, extant literature on XAI poses the issue that this might not be inherently true for AI models in general (Arrieta et al., 2020; Gilpin et al., 2018). Thus, our findings answer this issue and therefore, extend the existing body of knowledge for rule-based XAI approaches in text analytics.

2) *Global reconstruction by means of linguistic rules paves the way for a thorough understanding of language models.*

In contrast to the existing body of knowledge from local reconstruction approaches, the proposed approach based on linguistic rules enables a global reconstruction of language models (cf. Section "Discussion of the results"). Hence, linguistic rules constitute the first necessary step for global and thorough understanding of these black boxes, which cannot be achieved by local reconstruction approaches (cf. Section "Introduction"). With linguistic rules as vital instrument, researchers in the field of XAI can now not only focus on how to justify predictions of language models for text analytics tasks (e.g., by leveraging tests of statistical significance for linguistic rules in a global reconstruction for language model predictions), but also improve language models based on their understanding. In particular, our approach could be used to specifically reconstruct and analyze false predictions of language models, to detect its flaws and thereby, to enhance these language models. Furthermore, an analysis of linguistic rules reconstructing a language model's predictions could enable to derive deeper insights regarding effects of different types of review texts (e.g., reviews for search goods vs. experience goods or reviews of different consumer segments). That is, such analyses could support to analyze whether language model predictions for reviews of different review types vary in the NLP building blocks contained in the rules for global reconstruction. In particular, our approach allows for assessing the contribution of specific NLP building blocks to global reconstructions of language model predictions, which supports in enhancing the understanding of language model predictions in text analytics.

3) *Global reconstructions help to understand language model-detected features used for text analytics research*.

Our work also has implications for other research strands such as text analytics of online consumer reviews regarding star ratings (e.g., Binder, Heinrich, Klier, Obermeier, & Schiller, 2019; Goeken, Tsekouras, Heimbach, & Gutt, 2020) or review helpfulness (e.g., Yin, Bond, & Zhang, 2014). Here, many IS researchers aim at analyzing and explaining the relations between (aspect-based) sentiments and a target variable (e.g., star ratings or review helpfulness). To enable such analyses, it is necessary to extract high-quality features from large review datasets by means of state-of-the-art language models in a first step. Similar as in the practical application scenarios AS1-4), it is also vital for researchers to base their analyses and insights on reliable and comprehensible features. Hence, the comprehensible global reconstructions of language model predictions detecting such features may further enable a better understanding of the target variable based on the review texts as it reduces the opacity of the feature detection

in the first step in such analyses of online consumer reviews. That is, our approach can help to shed light into black-box language models used for feature extraction in IS text analytics research.

Implications for practice

1) *Global reconstructions with high comprehensibility can improve acceptance of language models, and support their adoption in practice.*

The language model BERT is already used in various applications (cf. AS1-AS4 in Section "Introduction"). Here, reconstructions with higher comprehensibility by means of our approach can help to shed light on these language models, and thereby, to improve acceptance of such models. In particular, a reconstruction by our approach allows to verify that a language model applied in an electronic marketplace does not discriminate against specific groups. For instance, when online customer reviews are analyzed (cf. AS1), our approach can confirm that specific groups of customers are not discriminated against (e.g., by assigning a negative sentiment to certain countries, ethnicities or genders). In text analytics-assisted recruitment processes (cf. AS4), the rules provided by the presented approach can be examined whether they contain arguments regarding gender or other discriminating attributes (detected by particular synsets) indicating undesired biases or discriminations. Similarly, our approach helps to reconstruct and justify BERT's predictions in chatbots (AS2) and finance applications (AS3). Further, the rules provided by our approach allow for algorithmic auditing based on the GDPR and thus to comply with regulatory requirements (e.g., the data controller has to be able to show that the data processing is fair according to the GDPR, which can be supported by analyses with respect to discriminations as outlined above). This is especially relevant since algorithmic auditing will likely become the gold standard for companies deploying AI models (Casey et al., 2019).

2) *Linguistic rules enable different relevant setups with respect to the trade-off between fidelity and comprehensibility depending on the requirements of different stakeholders for XAI approaches in practice.*

Our approach based on linguistic rules is particularly promising, as it enables to establish different setups with respect to the trade-off between fidelity and comprehensibility, allowing for more profound analyses (Gilpin et al., 2018). That is, reconstructions with higher fidelity might be leveraged by data scientists to analyze language model predictions in detail (e.g., in the context of algorithmic auditing, cf. Section "Introduction"). In addition, domain experts might leverage reconstructions with higher comprehensibility to assess the justifications (e.g., of BERT's aspect term detection) in a given domain. In particular, AI text analytics models in practice can thus be analyzed with different global reconstructions by means of our approach, which can be combined to gain more robust insights and to comply with regulatory requirements.

Conclusion

Global reconstruction of language model predictions such as for the state-of-the-art model BERT is an important issue in both research and practice, since it can enable to justify decisions based thereon in many application scenarios (e.g., in eCommerce or finance) and thereby allow to comply with necessary algorithmic auditing. In this paper, we thus proposed a global XAI approach in text analytics for reconstructing predictions of language models by linguistic rules. Further, we discussed the trade-off between fidelity and comprehensibility for the global reconstructions. For the analysis of our approach and the trade-off, we considered aspect term and sentiment term detection in two datasets of different domains regarding laptops as search goods and restaurants as experience goods. The results showed that linguistic rules enable global reconstructions of higher fidelity for language models, which paves the way for a thorough understanding of language models in the future. Further, our approach helps to understand language model-detected features used for further analytics in research. For practical application scenarios such as eCommerce, finance or online recruitment, our approach can improve acceptance of language models and thus support their adoption in text analytics.

Nevertheless, our research has some limitations, which could be starting points for future works. In this paper, we focused on the predictions of BERT without further considering the correctness of these predictions. Thus, our research could also be transferred to an analysis of BERT's prediction errors aiming towards a further enhancement of language models (i.e., by using linguistic rules to specifically reconstruct false predictions). Moreover, as we focused on the tasks of aspect and sentiment detection for search and experience goods in eCommerce, other NLP tasks in different domains would be possible for examination and could further substantiate our findings. In particular, our approach shall serve as a basis for further research that analyzes by means of statistical tests how BERT arrives at its predictions. Here, our work provides the necessary first step toward such insights.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . others (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Askira-Gelman, I. (1998). Knowledge discovery: comprehensibility of the results. In *Proceedings of the thirty-first Hawaii international conference on system sciences.* Symposium conducted at the meeting of IEEE.

Augasta, M. G., & Kathirvalavakumar, T. (2012). Rule extraction from neural networks – A comparative study. In *PRIME.* Symposium conducted at the meeting of IEEE.

Binder, M., Heinrich, B., Klier, M., Obermeier, A. A., & Schiller, A. (2019). Explaining the stars: Aspect-based sentiment analysis of online customer reviews. *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking Explainable Machines: The GDPR's' Right to Explanation'Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Tech. LJ*, *34*, 143.

Chatterjee, S. (2019). Explaining customer ratings and recommendations by combining qualitative and quantitative user generated contents. *Decision Support Systems*, *119*, 14–22.

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

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and measuring the geometry of bert. In *33rd Conference on NeurIPS*.

Coheur, L. (2020). From Eliza to Siri and Beyond. In *Proceedings of the IPMU 2020* (pp. 29–41).

Dai, H., & Song, Y. (2019). Neural aspect and opinion term extraction with mined rules as weak supervision. In *57th Annual Meeting of the ACL* (pp. 5268–5277). ACL.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. *ArXiv Preprint ArXiv:2010.00711*.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 NAACL* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. In *CoRR*. abs/1710.00794.

Fellbaum, C. (2013). Wordnet. In *The encyclopedia of applied linguistics*. Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1285

Förster, M., Hühn, P., Klier, M., & Kluge, K. (2021). Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations. In *Proceedings of the 54th Hawaii International Conference on System Sciences*.

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020a). Evaluating Explainable Artifical Intelligence-What Users Really Appreciate. In *Proceedings of the 28th European Conference on Information Systems (ECIS)*.

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020b). Fostering Human Agency: A Process for the Design of User-Centric XAI Systems.

Fortune Business Insights (2021). Natural Language Processing (NLP) Market Size, Share and Covid-19 Impact Analysis. Retrieved from https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933

Geng, Z., Zhang, Y. [Yanhui], & Han, Y. (2021). Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, *429*, 132–140. https://doi.org/10.1016/j.neucom.2020.12.037

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA).* Symposium conducted at the meeting of IEEE.

Goeken, T., Tsekouras, D., Heimbach, I., & Gutt, D. (2020). The Rise of Robo-Reviews-The Effects of Chatbot-Mediated Review Elicitation on Review Valence. In *ECIS*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1–42.

Heidari, M., & Rafatirad, S. (2020). Semantic Convolutional Neural Network model for Safe Business Investment by Using BERT. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1–6). IEEE. https://doi.org/10.1109/SNAMS52053.2020.9336575

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., & Szubartowicz, M. (2019). Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems. *Electronic Markets*, 1–21.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Conference of the North American Chapter of the ACL* (pp. 4129–4138). ACL.

Jumelet, J., & Hupkes, D. (2018). Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items. In *EMNLP Workshop BlackboxNLP* (pp. 222–231). ACL.

Kamps, J., Marx, M., Mokken, R. J., & Rijke, M. d. (2004). Using WordNet to measure semantic orientations of adjectives. In *4th LREC*. ACL.

Kim, N., Patel, R., Poliak, A., Wang, A., & others (2019). Probing what different NLP tasks teach machines about function word comprehension. In **SEM*. ACL.

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the Dark Secrets of BERT. In *EMNLP-IJCNLP* (pp. 4365–4374). ACL. https://doi.org/10.18653/v1/D19-1445

Liu, Q., Gao, Z., Liu, B., & Zhang, Y. [Yuanlin] (2015). Automated rule selection for aspect extraction in opinion mining. In *24th IJCAI*. AAAI.

Luo, B., Lau, R. Y. K., Li, C., & Si, Y.-W. (2022). A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1). https://doi.org/10.1002/widm.1434

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL System Demonstrations* (pp. 55–60). ACL. Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

O'Donovan, J., Wagner, H. F., & Zeume, S. (2019). The Value of Offshore Secrets: Evidence from the Panama Papers. *The Review of Financial Studies*, *32*(11), 4117–4155. https://doi.org/10.1093/rfs/hhz017

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 NAACL* (pp. 2227–2237).

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. (2018). Dissecting contextual word embeddings: Architecture and representation. In *EMNLP Workshop BlackboxNLP*. ACL.

Potnis, A. (2018). Illuminating Insight for Unstructured Data at Scale. Retrieved from https://www.ibm.com/downloads/cas/Z2ZBAY6R

Qi, P., Zhang, Y. [Yuhao], Zhang, Y. [Yuhui], Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *ACL System Demonstrations* (pp. 101–108). ACL. Retrieved from https://arxiv.org/pdf/2003.07082

Ramon, Y., Martens, D., Evgeniou, T., & Praet, S. (2020). Metafeatures-based Rule-Extraction for Classifiers on Behavioral and Textual Data. *ArXiv Preprint ArXiv:2003.04792*.

Repke, T., & Krestel, R. (2021). Extraction and Representation of Financial Entities from Text. In S. Consoli, D. Reforgiato Recupero, & M. Saisana (Eds.), *Springer eBook Collection. Data science for economics and finance: Methodologies and applications* (pp. 241–263). Cham, Switzerland: Springer k. https://doi.org/10.1007/978-3-030-66891-4_11

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *32nd AAAI Conference on Artificial Intelligence*.

Schiller, A. (2019). Knowledge Discovery from CVs: A Topic Modeling Procedure. In *Proceedings of the 14th International Conference on business informatics*.

Shrestha, Y. R., Krishna, V., & Krogh, G. von (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, *123*, 588–603. https://doi.org/10.1016/j.jbusres.2020.09.068

Steur, A. J., Fritzsche, F., & Seiter, M. (2022). It's all about the text: An experimental investigation of inconsistent reviews on restaurant booking platforms. *Electronic Markets*, 1–34.

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *Conference of the North American Chapter of the ACL* (pp. 380–385). ACL. https://doi.org/10.18653/v1/N19-1035

Sushil, M., Šuster, S., & Daelemans, W. (2018). Rule induction for global explanation of trained models. In *EMNLP Workshop BlackboxNLP* (pp. 82–97). ACL.

Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Nature Scientific Reports*, *11*(1), 1–13.

Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the ACL*.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., . . . others (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of the 2019 ICLR*.

Van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How does BERT answer questions? A layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Vilone, G., & Longo, L. (2021). A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model Agnostic Methods. *Frontiers in Artificial Intelligence*, *4*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP Workshop BlackboxNLP* (pp. 353–355). ACL. https://doi.org/10.18653/v1/W18-5446

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Conference of the North American Chapter of the ACL* (pp. 2324–2335). ACL. https://doi.org/10.18653/v1/N19-1242

Xu, S., Barbosa, S. E., & Hong, D. (2020). BERT Feature Based Model for Predicting the Helpfulness Scores of Online Customers Reviews. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Advances in Intelligent Systems and Computing. Advances in Information and Communication* (Vol. 1130, pp. 270–281). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-39442-4_21

Yang, Y., Uy, M. C. S., & Huang, A. (2020, June 15). *FinBERT: A Pretrained Language Model for Financial Communications*. Retrieved from https://arxiv.org/pdf/2006.08097

Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews. *MIS Quarterly*, *38*(2), 539–560. https://doi.org/10.25300/MISQ/2014/38.2.10

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. In *Proceedings of the IEEE/CIS 2018* (Vol. 13, pp. 55–75).

Zhang, R., Yang, W., Lin, L., & others (2020). Rapid Adaptation of BERT for Information Extraction on Domain-Specific Business Documents. *ArXiv Preprint ArXiv:2002.01861*.

5 Conclusion

"There are only two ways to live your life.
One is as though nothing is a miracle.
The other is as though everything is a miracle."
Albert Einstein (*1879; †1955)

5.1 Major Findings

The epoch-making and ever faster technological progress provokes disruptive changes and poses pivotal challenges for individuals and organizations. In particular, AI is a disruptive technology that offers tremendous potential for many fields such as information systems and electronic commerce. Therefore, this dissertation contributes to AI for online review platforms aiming at enabling the future for consumers, businesses and platforms by unveiling the potential of AI. To achieve this goal, the dissertation investigates six major research questions embedded in the triad of data understanding of online consumer reviews, enhanced approaches in RS and ABSA and explanations for RS and ABSA. The dissertation addresses these research questions and thereby, extends the existing body of knowledge by deriving new insights on online consumer reviews and their impact on online review platforms as well as by providing new AI approaches in RS and ABSA. The findings of each individual paper contained in this dissertation are given in the corresponding Sections 2, 3 and 4. Therefore, this section outlines the summarized major findings of the dissertation regarding new insights and new approaches.

On the one hand, the dissertation derives new insights regarding long-term rating dynamics (cf. Section 2.1), feature perspectives influencing consumer ratings (cf. Section 2.2), the impact of input data quality on RS (cf. Section 4.1) and the comprehensibility of language models (cf. Section 4.2). First, novel long-term rating trends and lasting effects of initial rating dynamics are found. In particular, a second disconfirmation of consumers in the long term of item ratings is discovered, which is indicated by an increasing rating trend in the long term that reverses the decreasing rating trend in the short term related to the first disconfirmation of consumers. Moreover, it is revealed that strong initial rating dynamics, based on initially stimulated consumer expectations, have a significant negative lasting effect on long-term average ratings. Second, new insights regarding feature perspectives influencing consumer ratings are found. In particular, it is revealed that the feature perspective user characteristics (e.g., consumer personality) has a significant and major contribution for explaining online consumer ratings. Third, a significant impact of input data quality on RS is established. While recommendations of RS show to be significantly more accurate overall when more features and feature values (i.e., completeness is increased) are available, it is also discovered that the impact of completeness on prediction accuracy is positively moderated by the amount of increased completeness per items and per consumers but negatively moderated by the diversity of added features. Fourth, it is shown that

language model's predictions for NLP tasks in online consumer reviews can be reconstructed globally and comprehensible by means of linguistic rules.

On the other hand, the dissertation provides new approaches for extending item content datasets in the context of RS (cf. Section 3.1), for enhancing MIL for ABSA with partly fine-grained supervision (cf. Section 3.2) and for global reconstructions of deep-learning AI models in NLP (cf. Section 4.2). First, a procedure is proposed for extending item content datasets in the context of RS. The procedure increases the completeness of a dataset by adding item content data from external data sources and by missing value imputation and thus, enables improved recommendations. Second, an enhanced approach for ABSA combining MIL and partly fine-grained supervision is established. The approach combines the advantage of leveraging high volumes of review-level rating data in online consumer reviews and of reduced labeling efforts for fine-grained supervision and thus, enables leading-edge results for ABSA in a very economical way. Third, an XAI approach based on linguistic rules for global reconstruction of language models is developed. The approach enables a comprehensible and global reconstruction of language models' predictions in text analytics tasks and allows for balanced setups with respect to the trade-off between comprehensibility and fidelity of the reconstruction.

5.2 Summary of Implications

The findings of this dissertation have various implications for research and practice. The detailed implications for all findings of the dissertation are contained in the individual papers in Sections 2, 3 and 4. Therefore, this section outlines a summary of these implications focusing on the major findings regarding new derived insights on online consumer reviews, RS and ABSA as well as regarding new approaches for enhancing and explaining AI approaches leveraging online consumer reviews.

By the derived insights on long-term rating dynamics (cf. Section 2.1) and feature perspectives influencing consumer ratings (cf. Section 2.2), the dissertation poses a first important step towards a systematic and thorough understanding of online consumer reviews and thus, towards a comprehensive understanding of consumers and their item assessments in electronic commerce. Furthermore, these insights are vital for businesses and platforms to improve marketing campaigns, product development and to improve AI approaches based on review data. In particular, the findings regarding long-term rating dynamics yield that especially sequential rating dynamics have to be considered by researchers, businesses and platforms when striving for representative and accurate consumer assessments in reviews, as ratings are influenced by two disconfirmations depending on the sequential rating order. Further, the findings also indicate, that consumers should be aware of the negative impact of strong initial rating dynamics on future ratings when relying on reviews for purchase decisions and not focus too strongly on initial one-sided (subjective) item ratings. Moreover, the findings and insights regarding important feature perspectives (e.g., consumer personality) influencing consumer ratings allow for building meaningful and multi-facetted item summarizations for online review platforms, for using relevant features of comprehensible perspectives for explaining recommendations to consumers as well as for improving items through the consumer criticism regarding important perspectives. In addition, the new findings regarding the positive impact of completeness of item content data on the

performance of RS (cf. Section 4.1) yield that increasing the quality of data, which RS operate on, can help to alleviate the stagnation in the field of RS. In particular, businesses can enable better recommendations for their items by providing additional data to online review platforms and the acquisition of additional data is advantageous for online review platforms, as improved recommendations enhance the efficacy of the platform. Moreover, the findings regarding global reconstructions of language models (cf. Section 4.2) show that linguistic rules enable global reconstructions of high fidelity for language models. This extends the existing body of knowledge in XAI for text analytics, as it was unclear whether comprehensible rule-based XAI models can exhibit high fidelity for reconstructions of AI models.

The proposed new approach for extending item content datasets in the context of RS (cf. Section 3.1) constitutes a tangible procedure in practical application scenarios which enables to increase data completeness with the aim of improving recommendation quality. In particular, this approach can significantly improve the recommendations for consumers, which strengthens consumer loyalty to platforms and businesses. In addition, this approach can serve as a template for researchers striving for similar investigations regarding the impact of other data quality dimensions (e.g., consistency or currency) on RS. Further, the enhanced MIL approach for ABSA with partly supervision (cf. Section 3.2) enables businesses as well as review platforms to conduct fine-grained sentiment analysis with good performance for online consumer reviews in a very resource-saving way. This is especially important in vast changing market environments, as in such cases, it would require high efforts to generate large datasets with labeled fine-grained sentiments every time an important change occurs (e.g., when new products or trends evolve on the market). Thus, this partly supervised MIL approach paves the way for an economical fine-grained sentiment analysis with high performance. In addition, the efficacy of this approach could encourage researchers in other fields to develop approaches combining MIL and supervised learning to leverage the advantages of both techniques. Moreover, the novel approach for rule-based global reconstructions of deep-learning AI models for text analytics (cf. Section 4.2) helps to improve the acceptance and trust in such black-box AI models and thus, supports their adoption in research and practice. This is especially important, since language models such as BERT are already used in various applications in electronic commerce and finance. In particular, to comply with regulatory requirements such as GDPR, the global reconstructions provided by this approach allow for algorithmic auditing, which will likely become the gold standard for companies deploying AI models (Casey et al., 2019). Furthermore, the approach is particularly promising, as it enables to establish different setups with respect to the trade-off between fidelity and comprehensibility, allowing for more profound analyses for different applications. In addition, the approach helps to reduce the opacity of features, which are extracted from text by language models such as aspect-based sentiments, used for analyses regarding their impact on business critical figures such as review ratings or review helpfulness. That is, the approach can help to shed light into black-box language models used for feature extraction in information systems research and practice.

Overall, the findings of this dissertation allow for better understanding of consumers as well as enhanced AI applications in RS and ABSA and their adoption in electronic commerce. Thereby,

the dissertation unleashes and exploits AI's potential for online review platforms enabling its future in the field of electronic commerce.

5.3 Directions for Future Works

The dissertation investigates six major research questions addressing important issues embedded in the triad of data understanding of online consumer reviews, enhanced approaches in RS and ABSA and explanations for RS and ABSA for online review platforms. Nevertheless, because of the tremendous and highly versatile potential of AI for multiple applications in electronic commerce and beyond, many interesting challenges and vital opportunities exist for future works aiming to further exploit the potential of AI. In the following, selected opportunities and possible directions for future works regarding AI, in particular, for online review platforms, are outlined.

To further improve understanding of online consumer reviews, it would be fascinating to analyze temporal dynamics of fine-grained sentiment information contained in the review texts as well as to examine interdependencies between feature perspectives in review texts. In particular, the dissertation focused only on the rating part of reviews for analyzing temporal dynamics (cf. Section 2.1). Therefore, it would be interesting to analyze whether similar sequential and temporal dynamics also exist in the fine-grained and aspect-based sentiments contained in the textual part of online consumer reviews. This could lead to deeper insights also regarding the overall rating dynamics. Moreover, the dissertation only investigated the direct impact of feature perspectives on review ratings (cf. Section 2.2). Here, analyzing interdependent moderator effects between these feature perspectives, such as whether consumer characteristics would influence the impact of item characteristics on review ratings, would be highly attractive and could complement the findings of this dissertation.

Further, enhancing RS and ABSA approaches by incorporating additional data and insights seems highly promising. In the case of RS, the dissertation focused on improving the data quality of item content data in the dimension of completeness (cf. Section 3.1 and 4.1). Here, data quality could also be improved by means of other dimensions such as accuracy or currency and analyses regarding resulting impacts on the performance of RS would be interesting. Moreover, utilizing features extracted from online consumer reviews by means of language models (e.g., consumer personality or aspect-based sentiments) could be vital for further enhancing RS. Similarly, such features extracted from texts could also be used to improve data quality by filling up missing feature values to increase completeness or by updating outdated feature values to increase currency. In addition, the dissertation used prediction accuracy for assessing the performance of RS, which is the most important quality measure for RS. However, high accurate recommendations are not the only goal of RS. For example, it can be beneficial to introduce new items to consumers, which create serendipity for consumers. Therefore, it would be interesting to use other evaluation criteria (e.g., coverage or serendipity) for measuring the performance of enhanced RS, especially as recommendations in different application scenarios may aim at different objectives. In the case of ABSA, the dissertation focused on leveraging partly supervision for improving MIL approaches aiming at instance-level sentiment classification (cf. Section 3.2). In particular, the MIL approach used in this work assesses the importance of instance's sentiments (e.g., sentiment of a sentence) for the review-level rating independent of the other instances in the review. However, additional

information regarding the review structure and hierarchy (e.g., by means of Rhetorical Structure Theory parsers; cf. Hou et al., 2020) could be vital to consider, as some information in the review might be redundant as elaborations of other main sentiments. In addition, it would be interesting to analyze whether using fine-grained sentiment labels from different datasets and domains could also enhance MIL approaches for fine-grained sentiment classification.

For explanations regarding language models, this dissertation focused on reconstructing the predictions of language models for aspect term and sentiment term detection (cf. Section 4.2). Here, the correctness of the language model predictions had not been taken into account. Thus, research could also be extended to analyses of these false predictions aiming towards further enhancement of language models (i.e., by using linguistic rules to specifically reconstruct false predictions). Moreover, the proposed approach using linguistic rules could also be transferred to and evaluated for other relevant tasks of NLP, such as text summarization or text classification.

Furthermore, the analyses and evaluations of this dissertation mainly focused on online consumer reviews in the widely discussed domains of restaurants, movies and laptops in research on electronic commerce. Here, it would be highly interesting to conduct data understanding as well as analyses of enhanced approaches and of explanations in RS and ABSA on other datasets of different domains, even outside the field of electronic commerce. This could further substantiate and broaden the findings of this dissertation and it could also reveal some new interesting findings. In addition, the analyses and evaluations were conducted in offline settings on self-contained datasets. Especially for new AI approaches, it would be vital to conduct online experiments that enable to assess performance measures on new consumer feedback experiencing the new approaches and to identify challenges emerging when deploying these approaches to real business applications. Here, it would also be interesting to analyze the scalability of the approaches regarding computational resources and processing time. Furthermore, it would be attractive to analyze the impact of the proposed approaches and the consequences of the derived data insights on economic models in real-world business applications (e.g., on online review platforms).

Moreover, future changes in online review platforms are on the rise which will pose new additional challenges for AI applications in this field. In particular, as more and more users generate multimodal content (e.g., audio and video) on social media, it is likely that more and more online consumer reviews will also be equipped with such media formats in the near future (Chandrasekaran et al., 2021). In particular, processing such multimodal content of high volumes poses new challenges for successful AI applications in business applications (e.g., emotion recognition; cf. Saxena et al., 2020), which have to be addressed. Moreover, the interactivity of online review platforms with its consumers is increasing (Adjei et al., 2022). Here, platforms could enable consumers to assess reviews of other consumers not only as a whole object whether it was helpful or not. In contrast, it could be possible that consumers can specify, confirm or revoke individual parts of consumer reviews (e.g., fine-grained sentiments of single sentences). Such new fine-grained consumer interaction data could open up a plethora of new possibilities for consumer understanding and enhanced approaches in RS and ABSA. Furthermore, it is on the rise that personalization in online review platforms will incorporate the consumers' social networks for improved recommendations (Suhaim and Berri, 2021). Here, online review platforms could be completely personalized and designed specifically for individual groups of friends or colleagues.

As the rapid technological progress is forecast to continue in the future (Benbya et al., 2021), new general challenges for AI applications will emerge that have to be addressed. In particular, AI approaches may be biased due to unbalanced datasets or error-prone modeling. Therefore, researchers have already drawn attention to analyze bias and fairness regarding AI (Mehrabi et al., 2021). Furthermore, with new technological advancements in AI, trends in RS and ABSA show that conversational RS will become more and more popular, which allow for humanoid communication with humans based on technology such as ABSA. Here, it is likely that AI will establish itself as a personal advisor assisting in decision making in everyday private and professional life (Singh et al., 2021). Therefore, it is becoming more and more urgent to fully understand AI, as more and more decisions will be based on AI.

To conclude, this dissertation exploits the potential of AI for online review platforms by contributing new insights and approaches regarding data understanding of online consumer reviews, enhanced approaches in RS and ABSA and explanations for RS and ABSA. Because of the tremendous and highly versatile potential of AI, many vital topics and critical issues still remain unanswered and thus, require future research. By means of its major findings and resulting implications, the dissertation opens up many new opportunities and challenges and paves the way for multiple interesting future works in this exciting area of AI in electronic commerce and beyond.

6 References

Abbasi, A., S. Sarker and R. Chiang (2016). "Big Data Research in Information Systems: Toward an Inclusive Research Agenda" *Journal of the Association for Information Systems* 17 (2), I–XXXII.

Adadi, A. and M. Berrada (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)" *IEEE Access* 6, 52138–52160.

Adjei, M. T., N. Zhang, R. Bagherzadeh, M. Farhang and A. Bhattarai (2022). "Enhancing consumer online reviews: the role of moral identity" *Journal of Research in Interactive Marketing*.

Angelidis, S. and M. Lapata (2018). "Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis" *Transactions of the Association for Computational Linguistics* 6, 17–31.

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

Baesens, B., R. Bapna, J. R. Marsden, J. Vanthienen and J. L. Zhao (2016). "Transformational Issues of Big Data and Analytics in Networked Business" *MIS Quarterly* 40 (4), 807–818.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *Information Fusion* 58, 82–115.

Baum, D. and M. Spann (2014). "The Interplay Between Online Consumer Reviews and Recommender Systems: An Experimental Analysis" *International Journal of Electronic Commerce* 19 (1), 129–162.

Bawack, R. E., S. F. Wamba, K. D. A. Carillo and S. Akter (2022). "Artificial intelligence in E-Commerce: a bibliometric study and literature review" *Electronic Markets*.

Bedué, P. and A. Fritzsche (2022). "Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption" *Journal of Enterprise Information Management* 35 (2), 530–549.

Benbya, H., T. H. Davenport and S. Pachidi (2020). "Special Issue Editorial. Artificial Intelligence in Organizations: Current State and Future Opportunities" *MIS Quarterly Executive* 19 (4).

Benbya, H., S. Pachidi and S. L. Jarvenpaa (2021). "Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research" *Journal of the Association for Information Systems* 22 (2), 281–303.

Binder, M., B. Heinrich, M. Hopf and A. Schiller (2022a). "Global Reconstruction of Language Models with Linguistic Rules – Explainable AI for Online Consumer Reviews" *Working Paper, University of Regensburg*.

Binder, M., B. Heinrich, M. Hopf and M. Szubartowicz (2022b). "The Way to the Stars: Explaining Star Ratings in Online Consumer Reviews" *Working Paper, University of Regensburg*.

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

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

Buck, A. (2022). *57 AMAZON STATISTICS TO KNOW IN 2022*. URL: https://landingcube.com/amazon-statistics/ (visited on 04/27/2022).

Casey, B., A. Farhangi and R. Vogl (2019). *Rethinking Explainable Machines: The GDPRs Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise*.

Chandrasekaran, G., T. N. Nguyen and J. Hemanth D. (2021). “Multimodal sentimental analysis for social media applications: A comprehensive review” *WIREs Data Mining and Knowledge Discovery* 11 (5).

Cramer-Flood, E. (2022). *Global Ecommerce Forecast 2022*. *As 2-Year Boom Subsides, Plenty of Bright Spots Remain*. URL: https://www.emarketer.com/content/global-ecommerce-forecast-2022 (visited on 05/06/2022).

Dacrema, M. F., S. Boglio, P. Cremonesi and D. Jannach (2021). “A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research” *ACM Transactions on Information Systems* 39 (2), 1–49.

Dang, C. N., M. N. Moreno-García and F. D. La Prieta (2021). “An Approach to Integrating Sentiment Analysis into Recommender Systems” *Sensors (Basel, Switzerland)* 21 (16).

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North*. Ed. by J. Burstein, C. Doran, T. Solorio. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 4171–4186.

Feng, J., X. Li and X. Zhang (2019). “Online Product Reviews-Triggered Dynamic Pricing: Theory and Evidence” *Information Systems Research* 30 (4), 1107–1123.

Firouzi, F., B. Farahani, M. Barzegari and M. Daneshmand (2022). “AI-Driven Data Monetization: The Other Face of Data in IoT-Based Smart and Connected Health” *IEEE Internet of Things Journal* 9 (8), 5581–5599.

Galov, N. (2022). *17+ Google Maps Statistics to Survey in 2022*. URL: https://webtribunal.net/blog/google-map-statistics/#gref (visited on 04/27/2022).

Ghasemaghaei, M. and G. Calic (2019). “Does big data enhance firm innovation competency? The mediating role of data-driven insights” *Journal of Business Research* 104, 69–84.

Godes, D. and J. C. Silva (2012). “Sequential and Temporal Dynamics of Online Opinion” *Marketing Science* 31 (3), 448–473.

Goldberg, L. R. (1990). “An alternative "description of personality": The Big-Five factor structure” *Journal of Personality and Social Psychology* 59 (6), 1216–1229.

Graham, M. and J. Elias (2021). *How Google’s $150 billion advertising business works*. URL: https://www.cnbc.com/2021/05/18/how-does-google-make-money-advertising-business-breakdown-.html (visited on 04/25/2022).

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi (2019). “A Survey of Methods for Explaining Black Box Models” *ACM Computing Surveys* 51 (5), 1–42.

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

Heinrich, B., T. Hollnberger, M. Hopf and A. Schiller (2022a). “Long-term Sequential and Temporal Dynamics in Online Consumer Ratings” *Proceedings of the 30th European Conference on Information Systems (ECIS 2022)*.

Heinrich, B., M. Hopf, D. Lohninger, A. Schiller and M. Szubartowicz (2021). “Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems” *Electronic Markets* 31 (2), 389–409.

Heinrich, B., M. Hopf, D. Lohninger, A. Schiller and M. Szubartowicz (2022b). “Something’s Missing? A Procedure for Extending Item Content Data Sets in the Context of Recommender Systems” *Information Systems Frontiers* 24 (1), 267–286.

Heinrich, B. and D. Hristova (2014). “A Fuzzy Metric for Currency in the Context of Big Data” *Proceedings of the 22nd European Conference on Information Systems (ECIS 2014)*.

Heinrich, B., D. Hristova, M. Klier, A. Schiller and M. Szubartowicz (2018a). “Requirements for Data Quality Metrics” *Journal of Data and Information Quality* 9 (2), 1–32.

Heinrich, B., M. Klier, A. A. Obermeier and A. Schiller (2018b). “Event-Driven Duplicate Detection: A Probability-based Approach” *Proceedings of the 26th European Conference on Information Systems (ECIS 2018)*.

Hoang, M., O. A. Bihorac and J. Rouces (2019). “Aspect-Based Sentiment Analysis using BERT”. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, pp. 187–196. URL: https://aclanthology.org/W19-6120.

Hopf, M. (2022). “Leveraging Fine-grained Supervision to Improve Multiple Instance Learning for Fine-grained Sentiment Classification in Online Consumer Reviews” *Working Paper, University of Regensburg*.

Hou, S., S. Zhang and C. Fei (2020). “Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications” *Expert Systems with Applications* 157, 113421.

Hu, H. and A. S. Krishen (2019). “When is enough, enough? Investigating product reviews and information overload from a consumer empowerment perspective” *Journal of Business Research* 100, 27–37.

Hu, N., P. A. Pavlou and J. Zhang (2017). “On Self-Selection Biases in Online Product Reviews” *MIS Quarterly* 41 (2), 449–471.

Jabr, W., Y. Chen, K. Zhao and S. Srivastava (2018). “What Are They Saying? A Methodology for Extracting Information from Online Reviews” *Proceedings of the 39th International Conference on Information Systems*.

Jabr, W., B. Liu, D. Yin and H. Zhang (2020). “Online Word-of-Mouth” *MIS Quarterly Research Curations*.

Jain, P. K., R. Pamula and G. Srivastava (2021). “A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews” *Computer Science Review* 41, 100413.

Jannach, D. and M. Jugovac (2019). “Measuring the Business Value of Recommender Systems” *ACM Transactions on Management Information Systems* 10 (4), 1–23.

Jannach, D., O. Sar Shalom and J. A. Konstan (2019). *Towards More Impactful Recommender Systems Research*.

Janssen, M., H. van der Voort and A. Wahyudi (2017). "Factors influencing big data decision-making quality" *Journal of Business Research* 70, 338–345.

Jesse, M. and D. Jannach (2021). "Digital nudging with recommender systems: Survey and future directions" *Computers in Human Behavior Reports* 3, 100052.

Karimi, M., D. Jannach and M. Jugovac (2018). "News recommender systems – Survey and roads ahead" *Information Processing & Management* 54 (6), 1203–1227.

Kaufer, S. (2022). *Tripadvisor, Inc., Form 10-K*. URL: https://ir.tripadvisor.com/static-files/3e32bcb6-bb03-47ea-bd68-e7a3e82c2d30 (visited on 04/27/2022).

Koren, Y. (2009). "The bellkor solution to the netflix grand prize" *Netflix prize documentation* 1 (81), 1–10.

Kotzias, D., M. Denil, N. de Freitas and P. Smyth (2015). "From Group to Individual Labels Using Deep Features". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Ed. by L. Cao, C. Zhang, T. Joachims, G. Webb, D. D. Margineantu, G. Williams. New York, NY, USA: ACM, pp. 597–606.

Kruse, L., N. Wunderlich and R. Beck (2019). "Artificial Intelligence for the Financial Services Industry: What Challenges Organizations to Succeed". In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Ed. by T. Bui: Hawaii International Conference on System Sciences.

Li, H. (2019). "Special Section Introduction: Artificial Intelligence and Advertising" *Journal of Advertising* 48 (4), 333–337.

Li, T., L. Lin, M. Choi, K. Fu, S. Gong and J. Wang (2018a). "YouTube AV 50K: An Annotated Corpus for Comments in Autonomous Vehicles". In: *iSAI-NLP 2018 proceedings*. *November 15-17, 2018, Pattaya, Thailand*. Piscataway, NJ: IEEE, pp. 1–5.

Li, X. and L. M. Hitt (2008). "Self-Selection and Information Role of Online Product Reviews" *Information Systems Research* 19 (4), 456–474.

Li, Y., Z. Zhang, Y. Peng, H. Yin and Q. Xu (2018b). "Matching user accounts based on user generated content across social networks" *Future Generation Computer Systems* 83, 104–115.

Liu, X., D. Lee and K. Srinivasan (2019). "Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning" *Journal of Marketing Research* 56 (6), 918–943.

Lu, Y. (2019). "Artificial intelligence: a survey on evolution, models, applications and future trends" *Journal of Management Analytics* 6 (1), 1–29.

Mauro, A. de, M. Greco and M. Grimaldi (2015). "What is big data? A consensual definition and a review of key research topics". In: AIP Publishing LLC, pp. 97–104.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman and A. Galstyan (2021). "A Survey on Bias and Fairness in Machine Learning" *ACM Computing Surveys* 54 (6), 1–35.

Milana, C. and A. Ashta (2021). "Artificial intelligence techniques in finance and financial markets: A survey of the literature" *Strategic Change* 30 (3), 189–209.

Mitra, P., C. A. Murthy and S. K. Pal (2002). "Unsupervised feature selection using feature similarity" *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (3), 301–312.

Moore, S. (2018). *How to Create a Business Case for Data Quality Improvement*. URL: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement (visited on 04/28/2022).

Müller, O., I. Junglas, S. Debortoli and J. vom Brocke (2016). "Using Text Analytics to Derive Customer Service Management Benefits from Unstructured Data" *MIS Quarterly Executive* 15 (4).

Musto, C., M. de Gemmis, G. Semeraro and P. Lops (2017). "A Multi-criteria Recommender System Exploiting Aspect-based Sentiment Analysis of Users' Reviews". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. Ed. by P. Cremonesi, F. Ricci, S. Berkovsky, A. Tuzhilin. New York, NY, USA: ACM, pp. 321–325.

Nguyen, N. (2018). *Inside Amazon's Fake Review Economy*. URL: https://www.buzzfeednews.com/article/nicolenguyen/amazon-fake-review-problem (visited on 04/27/2022).

Ntoutsi, E. and K. Stefanidis (2016). "Recommendations beyond the ratings matrix". In: *Proceedings of the Workshop on Data-Driven Innovation on the Web*. New York, NY, USA: ACM, pp. 1–5.

Ozsoy, M. G., F. Polat and R. Alhajj (2015). "Modeling Individuals and Making Recommendations Using Multiple Social Networks". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. Ed. by J. Pei, F. Silvestri, J. Tang. New York, NY, USA: ACM, pp. 1184–1191.

Pannala, N. U., C. P. Nawarathna, J. T. K. Jayakody, L. Rupasinghe and K. Krishnadeva (2016). "Supervised Learning Based Approach to Aspect Based Sentiment Analysis". In: *2016 IEEE International Conference on Computer and Information Technology (CIT)*: IEEE, pp. 662–666.

Panniello, U., M. Gorgoglione and A. Tuzhilin (2016). "Research Note—In CARSs We Trust: How Context-Aware Recommendations Affect Customers' Trust and Other Business Performance Measures of Recommender Systems" *Information Systems Research* 27 (1), 182–196.

Pappas, N. and A. Popescu-Belis (2017). "Explicit Document Modeling through Weighted Multiple-Instance Learning" *J. Artif. Int. Res.* 58 (1), 591–626.

Picault, J., M. Ribière, D. Bonnefoy and K. Mercer (2011). "How to Get the Recommender Out of the Lab?". In B. Shapira (ed.) *Recommender Systems Handbook*, pp. 333–365. New York, NY: Springer.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. de Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra and G. Eryiğit (2016). "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Ed. by S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, T. Zesch. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 19–30.

Repke, T. and R. Krestel (2021). "Extraction and Representation of Financial Entities from Text". In S. Consoli, D. Reforgiato Recupero and M. Saisana (eds.) *Data Science for Economics and Finance*, pp. 241–263. Cham: Springer International Publishing.

Ribeiro, M. T., S. Singh and C. Guestrin (2018). "Anchors: High-Precision Model-Agnostic Explanations" *Proceedings of the AAAI Conference on Artificial Intelligence* (32).

Ricci, F., L. Rokach and B. Shapira (2011). "Introduction to Recommender Systems Handbook". In B. Shapira (ed.) *Recommender Systems Handbook*, pp. 1–35. New York, NY: Springer.

Richthammer, C. and G. Pernul (2020). "Situation awareness for recommender systems" *Electronic Commerce Research* 20 (4), 783–806.

Rimol, M. (2021). *https://www.gartner.com/en/newsroom/press-releases/2021-11-22-Gartner Forecasts Worldwide Artificial Intelligence Software Market to Reach $62 Billion in 2022*. URL: https://www.gartner.com/en/newsroom/press-releases/2021-11-22-gartner-forecasts-worldwide-artificial-intelligence-software-market-to-reach-62-billion-in-2022 (visited on 04/25/2022).

Russel, S. J. and P. Norvig (2016). "Artificial intelligence: A modern approach" *Pearson Education*.

Sänger, J. and G. Pernul (2018). "Interactive Reputation Systems" *Business & Information Systems Engineering* 60 (4), 273–287.

Sänger, J., C. Richthammer, S. Hassan and G. Pernul (2014). "Trust and Big Data: A Roadmap for Research". In: *2014 25th International Workshop on Database and Expert Systems Applications*: IEEE, pp. 278–282.

Sar Shalom, O., S. Berkovsky, R. Ronen, E. Ziklik and A. Amihood (2015). "Data Quality Matters in Recommender Systems". In: *Proceedings of the 9th ACM Conference on Recommender Systems*. Ed. by H. Werthner, M. Zanker, J. Golbeck, G. Semeraro. New York, NY, USA: ACM, pp. 257–260.

Saxena, A., A. Khanna and D. Gupta (2020). "Emotion Recognition and Detection Methods: A Comprehensive Survey" *Journal of Artificial Intelligence and Systems* 2 (1), 53–79.

Shrestha, Y. R., V. Krishna and G. von Krogh (2021). "Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges" *Journal of Business Research* 123, 588–603.

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

Singh, P. K., P. K. D. Pramanik, A. K. Dey and P. Choudhury (2021). "Recommender systems: an overview, research trends, and future directions" *International Journal of Business and Systems Research* 15 (1), 14.

Song, X., S. Yang, Z. Huang and T. Huang (2019). "The Application of Artificial Intelligence in Electronic Commerce" *Journal of Physics: Conference Series* 1302 (3), 32030.

Statista (2020). *Revenues from the artificial intelligence (AI) software market worldwide from 2018 to 2025*. URL: https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues (visited on 04/24/2022).

Stumme, G. and A. Hotho (eds.) (2013). *Proceedings of the 24th ACM Conference on Hypertext and Social Media - HT '13.* New York, New York, USA: ACM Press.

Suhaim, A. B. and J. Berri (2021). "Context-Aware Recommender Systems for Social Networks: Review, Challenges and Opportunities" *IEEE Access* 9, 57440–57463.

Sun, Q., J. Niu, Z. Yao and H. Yan (2019). “Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level” *Engineering Applications of Artificial Intelligence* 81, 68–78.

Szczepański, M., M. Pawlicki, R. Kozik and M. Choraś (2021). “New explainability method for BERT-based model in fake news detection” *Scientific reports* 11 (1), 23705.

Tabakhi, S. and P. Moradi (2015). “Relevance–redundancy feature selection based on ant colony optimization” *Pattern Recognition* 48 (9), 2798–2811.

Vielberth, M., L. Englbrecht and G. Pernul (2021). “Improving data quality for human-as-a-security-sensor. A process driven quality improvement approach for user-provided incident information” *Information & Computer Security* 29 (2), 332–349.

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

Xu, H., B. Liu, L. Shu and P. Yu (2019). “BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis”. In: *Proceedings of the 2019 Conference of the North*. Ed. by J. Burstein, C. Doran, T. Solorio. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 2324–2335.

Xue, H.-J., X. Dai, J. Zhang, S. Huang and J. Chen (2017). “Deep Matrix Factorization Models for Recommender Systems”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Ed. by F. Bacchus, C. Sierra. California: International Joint Conferences on Artificial Intelligence Organization, pp. 3203–3209.

Yang, D., D. Zhang, Z. Yu and Z. Wang (2013). “A sentiment-enhanced personalized location recommendation system”. In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media - HT '13*. Ed. by G. Stumme, A. Hotho. New York, New York, USA: ACM Press, pp. 119–128.

Yi, C., Z. Jiang, X. Li and X. Lu (2019). “Leveraging User-Generated Content for Product Promotion: The Effects of Firm-Highlighted Reviews” *Information Systems Research* 30 (3), 711–725.

Yin, D., S. D. Bond and H. Zhang (2014). “Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews” *MIS Quarterly* 38 (2), 539–560.

Zhang, D., S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J. C. Niebles, M. Sellitto, Y. Shoham, J. Clark and R. Perrault (2021). *The AI Index 2021 Annual Report.* Stanford University.

Zhang, Q., L. Cao, C. Zhu, Z. Li and J. Sun (2018). “CoupledCF: Learning Explicit and Implicit User-item Couplings in Recommendation for Deep Collaborative Filtering”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Ed. by J. S. Rosenschein, J. Lang. California: International Joint Conferences on Artificial Intelligence Organization, pp. 3662–3668.