UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

Argumentative Explanations for Recommendations Based on Reviews

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Informatik und Angewandte Kognitionswissenschaft
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte kumulative Dissertation von

Diana Carolina Hernandez Bocanegra
aus Ibague, Kolumbien

1. Gutachter: Prof. Dr.-Ing. Jürgen Ziegler
2. Gutachter: Prof. Dr.-Ing. Torsten Zesch

Tag der mündlichen Prüfung: 30.03.2022

Abstract

Recommender systems (RS) assist users in making decisions on a wide range of tasks, while preventing them from being overwhelmed by enormous amounts of choices. RS prevalence is such that many users of information-based technologies interact with them on a daily basis. However, many of these systems are still perceived as black boxes by users, who often have no way of seeing or requesting the reasons why certain items are recommended, potentially leading to negative attitudes towards RS by users. Providing explanations in RS can bring several advantages for users' decision making and overall user experience. Although different explanatory approaches have been proposed so far, the general lack of user evaluation, and validation of concepts and implementations of explainable methods in RS, have left open many questions, related to how such explanations should be structured and presented. Also, while explanations in RS have so far been presented mostly in a static and non-interactive manner, limited work in explainable artificial intelligence have emerged addressing interactive explanations, enabling users to examine in detail system decisions. However, little is known about how interactive interfaces in RS should be conceptualized and designed, so that explanatory aims such as transparency and trust are met.

This dissertation investigates interactive, conversational explanations that enable users to freely explore explanatory content at will. Our work is grounded on RS explainable methods that exploit user reviews, and inspired by dialog models and formal argument structures. Following a user-centered approach, this dissertation proposes an interface design for explanations as interactive argumentation, which was empirically validated through different user studies. To this end, we implemented a RS able to provide explanations both through a graphical user interface (GUI) navigation and a natural language interface. The latter consists of a conversational agent for explainable RS, which supports conversation flows for different types of questions written by users in their own words. To this end, we formulated a model to facilitate the detection of the intent expressed by a user on a question, and collected and annotated a dataset helpful for intent detection, which can facilitate the development of explanatory dialog systems in RS.

The results reported in this dissertation indicate that providing interactive explanations through a conversation, i.e. an exchange of questions and answers between the user and the system, using both GUI-navigation or natural language conversation, can positively impact users evaluation of explanation quality and of the system, in terms of explanatory aims like transparency, and trust.

Keywords: Recommender systems, Explanations, Argumentation, Interactive interfaces design, Conversational agent, Dataset, Empirical studies.

Zusammenfassung

Empfehlungssysteme (Recommender Systems, RS) unterstützen die Nutzer bei der Entscheidungsfindung in einer Vielzahl von Aufgaben und verhindern gleichzeitig, dass sie von der enormen Menge an Auswahlmöglichkeiten überwältigt werden. RS sind so weit verbreitet, dass viele Nutzer von Informationstechnologien täglich mit ihnen interagieren. Allerdings werden viele dieser Systeme von den Nutzern immer noch als Blackboxen wahrgenommen, die oft keine Möglichkeit haben, die Gründe für die Empfehlung bestimmter Artikel zu sehen oder abzufragen. Dies kann zu einer negativen Einstellung der Nutzer gegenüber RS führen. Die Bereitstellung von Erklärungen in RS kann mehrere Vorteile für die Entscheidungsfindung der Nutzer und die allgemeine Nutzererfahrung mit sich bringen. Obwohl bisher verschiedene Erklärungsansätze vorgeschlagen wurden, hat der generelle Mangel an Nutzerevaluierung und die Validierung von Konzepten und Implementierungen erklärungsfähiger Methoden in der RS viele Fragen offen gelassen, die damit zusammenhängen, wie solche Erklärungen strukturiert und präsentiert werden sollten. Während Erklärungen in der RS bisher meist statisch und nicht interaktiv präsentiert wurden, gibt es nur wenige Arbeiten im Bereich der erklärbare künstliche Intelligenz, die sich mit interaktiven Erklärungen befassen und es den Benutzern ermöglichen, Systementscheidungen im Detail zu untersuchen.

Diese Dissertation untersucht interaktive, konversationelle Erklärungen, die es Nutzern ermöglichen, Erklärungsinhalte nach Belieben zu erkunden. Diese Dissertation basiert auf RS-Erklärungsmethoden, die Nutzerbewertungen verwerten, und ist von Dialogmodellen und formalen Argumentationsstrukturen inspiriert. Nach einem nutzerzentrierten Ansatz wird in dieser Dissertation ein Schnittstellendesign für Erklärungen als interaktive Argumentation vorgeschlagen, das durch verschiedene Nutzerstudien empirisch validiert wurde. Zu diesem Zweck haben wir ein RS implementiert, das Erklärungen sowohl über eine GUI-Navigation als auch über eine natürlichsprachliche Benutzungsschnittstelle liefern kann. Letztere besteht aus einem Konversationsagent für erklärbare RS, der Konversationsabläufe für verschiedene Arten von Fragen unterstützt, die von Benutzern in ihren eigenen Worten geschrieben werden. Zu diesem Zweck formulierten wir ein Modell, das die Erkennung der von einem Benutzer auf eine Frage ausgedrückten Absicht erleichtert, und sammelten einen Datensatz mit Textannotationen, der die Entwicklung von erklärenden Dialogsystemen in RS erleichtern kann.

Die Ergebnisse dieser Dissertation zeigen, dass die Bereitstellung interaktiver Erklärungen durch eine Konversation, d.h. einen Austausch von Fragen und Antworten zwischen dem Benutzer und dem System, sowohl durch GUI-Navigation als auch durch Konversation in natürlicher Sprache, die Bewertung der Erklärungsqualität und des

Systems durch die Benutzer positiv beeinflussen kann, und zwar in Bezug auf Erklärungsziele wie Transparenz und Vertrauen.

Schlüsselwörter: Empfehlungssysteme, Erklärungen, Argumentation, Entwurf interaktiver Schnittstellen, Conversational agent, Datensatz, empirische Studien.

Papers Contained in the Dissertation

- **Paper 1**: Hernandez-Bocanegra, D.C., Donkers, T., & Ziegler, J. (2020). Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*, 219-225. ACM, New York, NY, USA. doi: https://doi.org/10.1145/3386392.3399302
- **Paper 2**: Hernandez-Bocanegra, D.C., & Ziegler, J. (2020). Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics. *Journal of Interactive Media, i-com*, 19(3), 181–200. doi: https://doi.org/10.1515/icom-2020-0021
- **Paper 3**: Hernandez-Bocanegra, D.C., & Ziegler, J. (2021). Effects of Interactivity and Presentation on Review-Based Explanations for Recommendations. In: Ardito C. et al. (eds) Human-Computer Interaction – INTERACT 2021. INTERACT 2021. Lecture Notes in Computer Science, vol 12933. Springer, Cham. https://doi.org/10.1007/978-3-030-85616-8_35
- **Paper 4**: Hernandez-Bocanegra, D.C., & Ziegler, J. (2021). Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *3rd Conference on Conversational User Interfaces (CUI '21)*, 1-11. ACM, New York, NY, USA. doi: https://doi.org/10.1145/3469595.3469596
- **Paper 5**: Hernandez-Bocanegra, D.C. & Ziegler, J. (2021). ConvEx-DS: A dataset for conversational explanations in recommender systems. In *Proceedings of IntRS 21: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 1-18. url: http://ceur-ws.org/Vol-2948/paper1.pdf
- **Paper 6**: Hernandez-Bocanegra, D.C. & Ziegler, J. (2021). Explaining Recommendations Through Conversations - Argumentative Dialog Model and Comparison of Interaction Styles. Manuscript under review in ACM Journal Transactions on Interactive Intelligent Systems.

Acknowledgments

First, I would like to express my gratitude to my supervisor Prof. Dr.-Ing. Jürgen Ziegler, for his constant support and for having trusted me to be part of this project, which has contributed significantly to my professional and personal development. I also want to thank my second supervisor, Prof. Dr.-Ing. Torsten Zesch, for his support in regard to the natural language processing perspective of this work.

I also want to thank the members of the Interactive Systems group, as well as the members of the research training group "User-Centred Social Media", for their support and feedback on key aspects of this work. Furthermore, I would like to thank Marco Hellmann and Issam Hamo, for their special contributions to this work.

I want to thank Prof. Dr. Fernando Arango, Prof. Dr. Tobias H. Donner, Dr. Peter Murphy and Dr. Leigh Fernandez, for their trust in my work, and for their decision to guide me along this exciting path of the academy.

I want to thank my mother Patricia and my brothers Daniel and Carlos, who supported me in the idea of crossing the ocean to make my plans a reality. To my uncles Luis Alberto and Gonzalo, and aunts Angela and Nancy, as well as to los Hurtado, for always believing in me, and for all the love and support they have given us in difficult times. To Tati, for inspiring me to follow this path. And to my friends, Jeny, Mariluz and Erika, for their constant support, and for those weekly conversations that make life so much more joyful.

Last but not least, I want to thank Florian. Without his amazing support every day, this journey would have been much less enjoyable, I am sure.

Contents

List of Figures

1 Introduction

Recommender systems (RS) assist users in making decisions on a wide range of tasks, while preventing them from being overwhelmed by enormous amounts of choices. Based on inferences made about users' preferences, item ratings or comments reported by customers about items' features, RS aim to predict options that would be to the user's enjoyment and satisfaction. The prevalence of RS is such that active users of information technologies interact with this type of systems on a daily basis: RS define the events and news we see on social networks, the suggestions in online shops and streaming platforms, and the hotels and restaurants suggested in booking services. However, the way most RS work and how they arrive at final decisions remains largely opaque to users, which can lead to negative attitudes towards these systems.

On the other hand, explaining the recommendations issued by a RS has been shown to bring significant benefits for users, with respect to factors such as transparency (system explains how it works), effectiveness (system helps user to make better decisions), or trust (user has confidence in the system) (Tintarev & Masthoff, 2015). Many approaches to RS explanations utilize and reflect information on feedback reported by customers, or characteristics of recommended items, approaches related to popular RS methods, such as collaborative and content-based filtering. Explanations based on collaborative filtering inform that a recommendation is issued based on preferences of similar users or items that the user liked in the past, e.g. Amazon's "Customers who bought . . . also bought...", while content-based explanations present users with item features that can be relevant to them, e.g. He et al. (2015); Vig et al. (2009).

Alternatively, and promoted by recent advances in natural language processing (NLP), customer-generated reviews have become increasingly important as rich sources of information on the advantages and disadvantages of an item. Reviews often provide details about different aspects or features of an item, and express users' opinions or sentiments about it. The above can be useful not only for generating recommendations, but also for explaining them, leveraging customers' detailed reporting of items' performance. Existent explanatory RS methods based on reviews allow for the derivation of explanations such as "You might be interested in [feature], on which this product performs well" (Zhang et al., 2014), or graphical representations of pros and cons of the different aspects, using for example bar charts (Muhammad et al., 2016).

However, due to a general lack of user-centered evaluations and validation of concepts and implementations in explainable RS research, many questions remained open as to how best to present recommendations' rationale, so explainable aims such as transparency or trust in the system were met, bearing in mind that not all users would benefit from the same type of explanation. Furthermore, most approaches to explainable RS involve a static, single-step display of explanations, limiting users in scrutinizing system recommendations, in the event that they are not fully understood or accepted. On the other hand, providing interactive options for examining explanatory information at will might positively impact users' evaluation of RS. However, the way in which interactive explanatory interfaces for review-based RS should be designed remained elusive.

Thus, this dissertation proposes and investigates the concept of interactive, conversational explanations where users are free to explore explanatory information at will. To this end, this dissertation analyses explanations from the perspective of argumentation theories, particularly the class of models diverting from a static arrangement of assertions, considering instead a dialectical approach, and focusing on the exchange of arguments between two parties within a dialog (e.g. Walton (2011)). Grounding on such models, this work addresses explanations as a process of interactive argumentation, a conversation between the system and the user.

While research interest in conversational RS has increased in recent years (Jannach et al., 2020), providing explanations through a conversation between user and system is largely unexplored, so there is a lack of empirical evidence to support its potential benefit (Sokol & Flach, 2020). While conversational approaches are often associated with natural language interactions, such as those supported by chatbots, this dissertation adopt the interpretation by Jannach et al. (2020), who defines conversational explanations as the provision of explanatory information in an interactive, multi-turn, dialogical process that may be facilitated by a natural language conversation or GUI-based navigation.

This dissertation covers the comparison of different types of explanations that can be generated with review-based explanatory methods. It addresses the effect of explanation interface components and different display styles (e.g. textual, of graph-based), the effects and implications of providing explanations through interactive options, and the comparison of different interface types to present explanations (namely GUI-navigation and natural language conversation). To this aim, a user-centered iterative approach was followed, to 1) design explanation schemes under both static and interactive approaches; 2) design and implement a RS that provides explanations in a conversational manner; and 3) address users' evaluation of this dissertation proposal. This work involves the examination of how users interact with different types of explanatory interfaces, how they evaluate conversational explanations in RS (in terms of explanation quality, system transparency and effectiveness, and trust in the system), and how users characteristics such as decision making or visualization familiarity can influence such perception.

This dissertation aims to contribute to the understanding of how to provide better explanations in review-based RS, taking as example the domain of hotels, and ends with a set of guidelines in this regard, addressed to RS practitioners.

1.1 Description of the Problem

It is assumed that better explanations lead to better perception of the system by users, as it was earlier assumed that better RS performance could result in better user experience (Knijnenburg et al., 2012). While empirical evaluations with "real users on real systems" would be needed to confirm this assumption (Knijnenburg et al., 2012), the generalized lack of user-centered evaluation of RS explanatory approaches and methods (Nunes & Jannach, 2017) has prevented finding the optimal way to present explanations, so that explanatory objectives are actually met.

This dissertation focuses on RS explanatory methods that exploit customer reviews. These are, however, a very subjective source of information, with significant variations in reliability, the aspects described and the language used. Even if the hurdle of extracting relevant aspects and sentiments from the noisy review texts is overcome, the question of what review-based information to show and how to present it remains largely open. Many questions arise at this point. Do textual summaries of opinions work better as explanations, or aggregated views of opinions? Is it better to provide graphical representations of the information, or purely textual statements? What level of detail is appropriate for presenting explanatory information? What kind of such information is actually relevant to users?

Even if answers to these questions are found, a new concern arises: How to enable users to contest the system when the explanations have not been fully accepted or understood? In this regard, the many differences in the explanatory needs of users, as well as different personal habits to process information when making decisions, among many other factors, mean that there is no univocal way of presenting explanations that can satisfy all users. Thus, it seems reasonable to establish which interactivity mechanisms could facilitate a flexible access and exploration of explanatory information at different levels of detail, so that the user can reach a better understanding of the reasons behind a recommendation.

However, although some work in the field of explainable artificial intelligence (XAI) has already addressed interactive explanations, their effective design and impact on RS remains also largely unexplored, as does the empirical validation of their effects on users in relation to explanatory aims.

1.2 Background

1.2.1 Recommendation methods

Recommender systems (RS) assist users in the selection of alternatives, while sparing them from dealing with enormous amounts of options. Such systems aim to generate personalized recommendations on the basis of an inferred model of user preferences and item quality. RS methods solve the task of ranking prediction, that is, how a user would rate a given item, taking into account the users' profile. The latter could be inferred, in turn, from the user's previous interactions with the system, or from explicitly stated preferences. Most popular RS methods can be categorized under the following types:

Content-based filtering. Here, users' profile is calculated based on values of features of items that the user has purchased or expressed preferences for, in the past (see e.g. Lops et al. (2011)). This methods utilize features or characteristics that usually correspond to item's metadata. Under this approach, for example, a movie recommender can suggest the user to watch "Forrest Gump", assuming a preference for Tom Hanks' movies, after the user previously reported they liked "Cast Away", "Apollo 13" and "Saving Private Ryan". In this example, the item feature is the "main actor", and the value is "Tom Hanks".

Collaborative filtering. Here, predictions are based on detected patterns that items / users share in common. This approach is grounded on the idea that similar users (neighbors) would like similar items. Methods under this category make use of explicit ratings granted by users to items (see e.g. Ekstrand et al. (2011)). Under this approach, a movie recommender would suggest the user from the previous example to watch "Schindler's list", given that most of the users who liked "Saving Private Ryan" also liked "Schindler's list".

A special case within this category of methods, are the latent factor models, including algorithms of the matrix factorization type. Here, similarities between items/users are calculated on the basis of latent representations, that is, underlying patterns of similar items/users are detected, rather than using only explicit ratings or purchasing history (as in plain collaborative filtering), or explicit feature values (as in content-based filtering). This dissertation focuses on RS under the matrix factorization approach, particularly those that integrate user reviews as an alternative source of information, as elaborated later in this section.

Hybrid methods. Under this approach are methods that combine content-based and collaborative filtering strategies (see e.g. Burke (2007)). Here, the system would suggest the user from previous examples, to watch "Interstellar", on the basis that, not only most of the users who liked "Apollo 13" also liked "Interstellar", but also share the same theme, i.e. "space travel movies".

1.2.2 Explainable RS

Explanations in RS seek to provide reasons behind a recommendation, while assisting users making a decision. By providing explanations, the evaluation of RS by users can be improved, in relation to aspects such as system transparency or trust in the system (Tintarev, 2007).

Explanation interface components. According to the taxonomy by Nunes & Jannach (2017), explanations in RS involve usually the following types information, or user interface components:

- Input parameters: refer to the input used to reach a decision or recommendation, e.g. in a music recommender, musical genres listened to in the last few days.
- Knowledge base (background or user knowledge): the explanations may reflect the item alternatives and features, as well as the matching between the recommended items and the users' preferences.
- Decision inference process (data or rationale of the inference method): the RS may provide indications on the recommendation process, or on the data used for it, e.g. "We suggest X because similar users liked it". Procedural and algorithmic explanations involve this kind of component.
- Decision output: focus on the decision outcome, for example items' quality in the form of pros and cons.

To date, a number of explanatory methods have been proposed in RS, which make it possible to provide explanations involving one or more of these components. However, it remained unclear the extent to which each component impacts users' evaluation of RS, or whether a particular component can benefit users more than others, depending on the context, as discussed by Nunes & Jannach (2017). Questions raised in this regard are for example, whether users benefit (and are even interested) in getting algorithmic details in non-critical contexts, as in hotel booking; or how useful it is for users to get details on their preferences inference, compared to explanations focused on the quality of the items.

Thus, this dissertation contributes with answers to questions in this respect, addressing the design of explainable interfaces involving the four types of components categorized by Nunes & Jannach (2017), as well as their individual impact on the evaluation of RS by users.

Explainable RS methods and display styles. Zhang & Chen (2020) suggest a taxonomy to classify explainable methods, based on the dimensions: type of model (which reflects the underlying RS method), and the information and style of the explanations, as summarized bellow.

- Type of model. Among the most popular approaches are the methods based on collaborative filtering, which allow to generate explanations based on relevant users or items, in a nearest-neighbor style (e.g. "Your neighbors' ratings for this movie" Herlocker et al. (2000)), as well as the content-based methods, that allow personalized, feature-based explanations to be generated providing users with indicators, such as the relevance of item features and how they match their preferences (e.g. Vig et al. (2009)).
- Information and style of the explanations. Popular approaches involve informing on: relevant users or items (as facilitated by the collaborative filtering methods); item/user features (as facilitated e.g. by content-based approaches); and customer opinions about items (e.g. pros and cons reported in reviews). Popular methods allow the generation of: textual explanations; visual explanations (using images or graphs); and social explanations (reflecting e.g. friends who like an item).

Explanations in RS could also be classified by their presentation format as defined by Nunes & Jannach (2017): natural language (e.g. canned text, template-based, structured language), visualization, or other media formats (e.g. audio).

The approach followed in this dissertation correspond to collaborative filtering explainable methods, particularly those based on matrix factorization algorithms. Previous work has addressed the problem of "how to explain" recommendations, addressing the comparison of different explanation styles based on collaborative filtering and content-based approaches - without involving user reviews - but leveraging explicit users' ratings and item features or tags. Herlocker et al. (2000) compared explanation styles using collaborative filtering techniques (using e.g. histograms, tables, aggregated numbers and text), and found that neighbor-style explanations ("your neighbors' ratings for this movie:") through histograms were rated as the most compelling. Bilgic & Mooney (2005) compared different explanations styles in a hybrid RS in the domain of books, and found that keyword-style explanations using a table were more beneficial to users in terms of effectiveness, compared to the neighbor-style bar charts. Gedikli et al. (2014)

extended the evaluation in Herlocker et al. (2000), by including further styles in their comparison (pie charts, tag clouds). Authors found that tag cloud explanations were beneficial to users, in terms of transparency and user satisfaction, despite their higher demand in cognitive effort by the users.

In contrast, while different explanation styles have been proposed for review-based explanation methods (see subsection below), to our knowledge, no comprehensive empirical comparison of such styles was reported, nor involving domains where decision making relies heavily on textual opinions expressed by customers, e.g. the hotel domain. Therefore, it remained elusive whether conclusions drawn from studies in Herlocker et al. (2000), Bilgic & Mooney (2005) and Gedikli et al. (2014) could generalize to our examined context; or which of the review-based styles described below could benefit users the most, in terms of explainable aims such as transparency or trust. Thus, this dissertation explores exhaustively the presentation possibilities enabled by review-based methods, comparing different types of textual and visual styles, as well as different types of user interface, to provide explanations reflecting opinions reported in customer reviews.

Explainable RS methods based on reviews. There has been increased interest in the use of user reviews in both RS and explainable RS, given the richness of information reported on diverse aspects, which cannot be deduced from the overall item ratings or content features. Among the explanatory methods based on reviews we have:

- Abstractive summarization methods based on natural language generation (NLG) techniques. In this case, the explanations can be presented as a verbal summary of the content found in reviews, without providing aggregated numbers or statistical information, but statements in natural language (e.g. Carenini et al. (2013); F. Costa et al. (2018)).
- A selection of helpful reviews or excerpts of them that might be relevant to the user, detected using deep learning techniques and attention mechanisms. Here, helpful reviews may enhance the accuracy of RS predictions, and be used and presented as explanations (e.g. C. Chen et al. (2018); Donkers et al. (2020)).
- An overview of the pros and cons regarding specific item features. Here, topic modelling and aspect-based sentiment analysis are usually used to detect the sentiment polarity towards item aspects or features (e.g. Dong et al. (2014); Wu & Ester (2015); Zhang et al. (2014)), information that is integrated to RS algorithms such as matrix or tensor factorization (e.g. Bauman et al. (2017); Wang et al. (2018); Zhang et al. (2014)). In this case, explanations can be presented using text templates (e.g. "You might be interested in [feature], on which this product performs well" (Zhang et al., 2014)), or providing an aggregated view of opinions, using

percentages or proportions in visual representations such as bar charts (Muhammad et al., 2016), or word clouds (Wu & Ester, 2015).

Providing verbal summaries and helpful reviews as explanations have proven to be an effective means of assisting users in making purchasing decisions, while helping them cope with the overwhelming amount of information available (Carenini et al., 2013; Ghose & Ipeirotis, 2011; Hu et al., 2017; Mudambi & Schuff, 2010; Pang & Lee, 2008). However, we argue that these explanation styles pose potential disadvantages: textual summarizations could be perceived as rather imprecise and subjective, since the evaluation of the quality of the item is generally presented based on adjectives such as "good", "great"; while a helpful review may be perceived as an anecdotal view, raising questions as to whether such a single report adequately represents the majority opinion. Also, providing such types of explanations may in turn result in additional explanation needs. A user might wonder, for example, what criteria the system has to establish that a certain review can be considered relevant to them. Thus, the explanation would in turn need an explanation.

A workaround to these drawbacks is to provide an aggregation of opinions with some kind of statistic, e.g. number or proportion of negative and positive opinions, by aspect, which may serve as easy anchors to convey more compelling information. In fact, judgments and decision making can be influenced by changes in attitude, which in turn can result from the effortless use of cues such as numerical anchors, when people lack motivation or ability during decision-making (Petty & Cacioppo, 1986; Wegener et al., 2010). Consequently, we focus on methods that facilitate such statistics and, in turn, integrate this type of information into ratings prediction. In this way, we could: 1) count on additional evidence to support the aggregated figures, i.e. the comments extracted from the reviews, to be displayed in case user requires such a level of detail; 2) provide explanations consistent with the recommendation algorithm, and not system generated post-hoc justifications, that may not be aligned with the diversity of opinions expressed by customers, nor the users' preferences. The Explicit Factor Method proposed by Zhang et al. (2014) meets these expectations, and it is introduced below.

Explicit Factor Method (EFM) (Zhang et al., 2014). This method exploits user-written reviews to generate recommendations and explanations. This approach lies under the collaborative filtering group of methods, more precisely a matrix factorization algorithm, which consist of a matrix representation of users' preferences and item qualities based on explicit ratings, and integrate them with additional matrices that reflect aspect-based sentiments expressed by customers in their reviews, e.g. "I think the hotel staff was great".

Matrix factorization models work on the basis of obtaining latent representations of characteristics that items may have in common, based on ratings granted to items by similar users. Latent features are in fact problematic in terms of explainability, since they are basically numerical expressions that do not represent tangible concepts as such, so they can hardly be explained in practical terms to end users. The EFM seeks to mitigate such limitation, by aligning the latent features with a (numeric) evaluation of explicit features extracted from reviews. Thus, besides the traditional rating matrix used in matrix factorization algorithms, two additional matrices are constructed: a user preference matrix, containing how many times a user addressed a feature in their reviews; and the item quality matrix, indicating how many positive / negative comments about a feature - were found in reviews for each item. Finally, an optimization task integrates such elements, to predict item ratings, which can be explained by means of explicit features information consolidated in the item quality and user preferences matrices.

By using the EFM, explanations of the type: "We recommend this hotel because 95% of customers have made positive comments about the staff, an aspect that is relevant for you" can be provided. However, EFM authors limited their user evaluation to providing brief explanations such as "You might be interested in [feature], on which this product performs well" and testing to what extent users examined further the recommended items, while the quality of the explanations was not evaluated, nor was the effect of such explanations on explanatory objectives such as transparency or trust. Thus, we extend the work proposed by Zhang et al. (2014), by using the EFM as basis of our implemented RS, but proposing and evaluating novel ways to provide explanatory information, leveraging such method.

Presentation styles. In regard to visualization techniques applied to review-based RS, Muhammad et al. (2016) proposed a summary of the positive and negative opinions on different aspects using bar charts, while Wu & Ester (2015) proposed the use of word clouds or radar charts to display such information.

While statistical explanatory information can be provided using visual displays (Muhammad et al., 2016), users with lower visual abilities might benefit less from a presentation based on images or graphics (Kirby et al., 1988; Schnotz, 2014), compared to a presentation using only text, and found that textual explanations were reported as more persuasive than the explanations provided using a visual format; however, users with greater visualization familiarity reported one of the visual format explanations more positively (a Venn diagram).

Nevertheless, to our knowledge and as discussed above, no comprehensive empirical comparison of these styles has been made. Thus, we examined in this dissertation (Paper 2 and Paper 3) the effects of different presentation styles for aggregated statistics of customer opinions on users evaluation of the RS and its explanations.

1.2.3 Interactive and conversational explanations

It has already been shown that interactive elements can improve the user experience in RS, mainly by providing control mechanisms over the criteria that influence the recommendation process itself (L. Chen & Pu, 2014; Loepp et al., 2015, 2014). However, little is known about the impact that explanations with interactive elements may have on the RS users' experience.

In this regard, the predominant approach in both RS and explainable AI (XAI) is to provide explanations in a static manner (i.e., using a non-interactive presentation) (Abdul et al., 2018), limiting users in scrutinizing system decisions when they are not fully understood or accepted. On the other hand, providing interactive options for examining explanatory information might positively impact users' evaluation of intelligent systems, by allowing the user to request, for example, further evidence for system claims and predictions.

For example, Krause et al. (2016) proposed a method to interactively visualize how specific features or data points affect machine learning predictions, and Sokol & Flach (2020) proposed a system supporting *why?* questions within an interactive dialog, to facilitate the understanding of ML classification outcomes. These approaches aim to meet the explanation needs of AI domain experts, and work on discrete and categorical data. On the other hand, our approach seeks to satisfy the explanatory needs of non-expert AI users, by exploiting subjective and unstructured information sources.

Conversational RS. Conversational approaches to RS may be realized in different ways, for example, as a natural language dialog with a conversational agent (e.g. chatbot), or through interaction steps in a graphical interface. In this sense, this dissertation adopts the interpretation of the term *conversational* by Jannach et al. (2020), which does not exclude forms of interaction outside written or spoken text. Thus, our proposal regards conversational explanations as the provision of explanatory information in an interactive, multi-turn, dialogical process that may be instantiated as natural language dialog or GUI-based navigation. Jannach et al. (2020) classify conversational approaches to RS into the following categories:

1) Conventional web-based navigation, based on structured layouts and features, like buttons and hyperlinks. This dissertation refers to this modality as *GUI navigation*, elaborates in Paper 3 the explanation approach under this paradigm, and compares it in Paper 6 to the natural language conversation approach.

2) Natural language, both written or spoken.

3) Hybrid, a combination of natural language and other modalities. Under this approach, users can, for example, indicate their input both by typing natural language expressions and using features such as buttons and other web controls. Paper 6 elaborates our approach to explanations under this paradigm, which is referred to in this dissertation as *natural language conversation*.

Most conversational RS focus on requesting user's preferences and recommending items, while little attention has been devoted to explaining decisions/predictions (Jannach et al., 2020). For example, work by Christakopoulou et al. (2016) and Zhang et al. (2018) proposes methods to elicit user preferences and to generate recommendations through dialog, while no explanations for system predictions are provided. Further work on conversational agents in the hotel domain usually focus on customer service and booking assistance Buhalis & Cheng (2020). In contrast, this dissertation explores the implications and effects of using conversational interfaces to explain recommendations.

Explanations as natural language conversation. Most interactive approaches in RS and, in a wider scope, in XAI, are based GUI navigation options. However, recent developments in natural language processing (NLP) and natural language generation (NLG) enable a more flexible interaction, where users could indicate, in their own words, their explanation needs.

This dissertation explore the feasibility and implications of using conversational agents to provide explanations in review-based RS, given their ability to enable two-way natural language communication, opening up the range of possible questions a user can ask the system. Although user interfaces inspired by human-to-human conversation have been developed and used for a long time to assist users in a wide range of tasks (Moore & Arar, 2018), little is known about how a conversational agent should be conceptualized or designed in the context of XAI, and in particular, in explainable RS.

In this regard, Rago et al. (2020) proposes a protocol for conversational explanations in RS, however it restricts the possible user interactions to a limited set of questions, while under our approach, users can indicate their explanatory needs using their own words. Further formal models for explanation as a dialog have been conceptually proposed as theoretical basis to the design of conversational explanation approaches (see e.g. Arioua & Croitoru (2015); Walton (2011)). However their practical implementation in RS (and in XAI in general) still lacks sufficient empirical evaluation (Madumal et al., 2019; Miller, 2018; Sokol & Flach, 2020), so it remained elusive how conversational interfaces should be designed in the context of RS, in order to improve the evaluation of the system by users. Thus, this dissertation contributes with an empirically validated interface design for explanations as interactive conversation, based on dialog models of explanation, as elaborated further in section 1.4.

Conversational agents. Dialog systems, often referred to as conversational agents, enable human-computer interaction by means of natural language statements, and can be categorized as non-task-oriented (e.g. smalltalk), or task-oriented (e.g. booking, or assistants such as Siri or Cortana) (H. Chen et al., 2017). Our approach corresponds to the text-based, task-oriented group.

The most prevalent approaches to task-oriented conversational agents can be grouped into pipeline and neural end-to-end methods (H. Chen et al., 2017). The pipeline approach is the most widely applied, and is characterized by architectures involving the components: 1) natural language understanding (to interpret the intent expressed in a user's dialog move), 2) dialog state tracker (to determine the dialog state based on input and dialog history), 3) dialog policy learning (to determine the next dialog action), and 4) natural language response generation. End-to-end methods, which have recently raised a growing interest, allow for joint training of all components, and can be beneficial in contexts where flexible adaptation is required, e.g. when the system is to be rapidly scaled to new domains or applications. However, such methods usually require very large datasets (of the order of thousands of dialogues) to be effective (Li et al., 2018).

Paper 6 elaborates the implementation process of our conversational agent for explainable review-based RS, which involves a pipeline architecture. Despite the advantages posed by neural end-to-end methods, and given the overall lack of datasets specifically focused on explanatory dialogues in RS (discussed in detail in Paper 4 and Paper 5), we opted for the pipeline approach. Adopting such architecture is beneficial during early stages of dialog engineering for new purposes (H. Chen et al., 2017), as in our case of explainable RS. Thus, we leave for future work the exploration of neural end-to-end approaches to conversation.

Intent detection and slot filling. Our implemented conversational agent is able to reply automatically to users' questions as part of an explanatory conversation. To this end, we set our focus on the natural language processing tasks that are key to the development of conversational agents: intent detection and slot filling. The former aims to interpret the user' information need expressed through a query, while the latter aims to detect which entities - and also features of an entity - the query refers to. The idea behind the *intent* concept is that user utterances within a dialog can be framed within a finite and more limited set of possible dialog acts (Verberne et al., 2013).

Methods proposed to solve the intent detection task range from conventional text classification methods, to more complex neural approaches, based on recurrent neural networks, attention-based mechanisms and transfer learning, to solve the intent detection and slot-filling tasks, both jointly and independently, and to extend the solutions to new domains, as surveyed by Louvan & Magnini (2020). The most common approach for intent representation, in the open-search domain, is intent classification (Verberne et al., 2013), that is, a query can be categorized according to a classification scheme, consisting of dimensions or categories, and their possible values (Broder, 2002; Verberne et al., 2013). This approach facilitates the implementation of automatic intent detection procedures, since detection can be solved by splitting a complex task into several text classification tasks, one per each dimension, for which methods based in language models like the state-of-art BERT (Devlin et al., 2019) can be leveraged. This dissertation adopts this approach, as discussed in detail in Paper 4 and Paper 5, where we propose a dimension-based intent model for intent detection in explainable RS, helpful to infer the explanation need expressed by the user in a query, as the combination of values detected for each dimension of the model.

Dialog management. Dialog state tracker and dialog policy learning are usually performed by a dialog management component. Harms et al. (2019) differentiates between two main approaches: handcrafted (state and policy are defined as a set of rules defined by developers and domain experts) and probabilistic (rules are learned from corpora with real conversations). Our implemented conversational agent corresponds to the handcrafted approach known as finite-state, where the dialog state has a fixed set of possible transitions to other states. The above given a lack of an existing corpus to train inference procedures for states and sequences for conversations in the explanatory RS context. Also, despite more sophisticated approaches such as the frame-based used in Google's DialogFlow [1] allow greater flexibility (e.g., adding a data model so that slots can be filled in any dialog sequence), these capabilities were not necessary for the purposes of this dissertation, so we leave their exploration for future work.

Question and answering systems. Work reported in this dissertation relates to question answering (QA) systems, which aim to answer user-written questions, by using information retrieval or natural language processing methods, on web documents or knowledge bases. However, in our examined context, explanations should not be generated purely from information sources, but should also reflect the mechanism used to generate the recommendations. Also, while most of QA systems respond to factoid questions (e.g. "does this hotel have a pool?"), much less work has been devoted to advanced "how-to", "why", evaluative, comparative, and opinion questions (Lim et al., 2009; Mishra & Jain, 2015), which are the type of questions usually asked in explanatory conversations. In addition, in contrast to the prevalent QA approach (system replies to standalone questions), interactive QA involves a dialog interface enabling related, follow-up and clarification questions (Quarteroni & Manandhar, 2008), an approach closer to ours, as elaborated further in Paper 4.

[1] https://cloud.google.com/dialogflow

1.2.4 Moderation effect of user characteristics

Besides explanations' content and interface design features, user characteristics may also contribute to differences in the RS users' experience. Knijnenburg et al. (2012) and Xiao & Benbasat (2007) argue that users' evaluation of the interaction with a RS usually depends on personal characteristics, such as demographics and domain knowledge. We then assumed that this would also be the case for the explanation quality, as addressed for example by Berkovsky et al. (2017) and Kouki et al. (2019). Particularly, Berkovsky et al. (2017) examined the moderation effect of users' personality traits on trust, given different types of explanations, in the movies domain, using to this end participants' scores of the Big-Five personality traits (openness, conscientiousness, extraversion, agreeableness and neuroticism) (P. T. Costa & McCrae, 1992; Tkalcic & Chen, 2015).

However, we opted to address user characteristics closer related to how users process information when making decisions, noting that supporting this process is precisely the goal of RS. Research on decision-making has shown that it is determined significantly by preferences and abilities to process available information (Driver et al., 1990). Hamilton et al. (2016) define two decision making styles: rational and intuitive, the former characterized by a propensity to search for information and evaluate alternatives exhaustively, and the latter by a quick processing based mostly on hunches and feelings. We therefore included this factor in our research to investigate its moderating effect.

Furthermore, since review-based explanations rely on the expressed opinions of other users, we also addressed effects of social awareness, i.e. of the extent to which users are inclined to adopt the perspective of others, when making decisions (Collaborative for Academic Social and Emotional Learning (2013) [CASEL]). The rationale for this interest stems from the tendency of individuals to adjust their own opinions using those of others, while choosing between various alternatives (Sniezek & Buckley, 1995), which may benefit their decisions (Yaniv & Milyavsky, 2007). Particularly, individuals with greater perspective-taking skills tend to understand the views of others better (Burack et al., 2006; Chandler, 1973), a trait defined as social awareness by CASEL (2013) .

A final factor we considered is the extent to which a user is familiar with graphical or tabular representations of information. Visualization familiarity may also influence user experience, when using images or graphs, as found by Kouki et al. (2019).

1.3 Overall Research Questions

To contribute to bridge the gaps described in the previous section, the aim of this dissertation is answer to the following overall questions, in regard to review-based RS:

- How can the quality of explaining recommendations be improved through interactive conversation?
- How does information extracted from user reviews need to be structured and presented, within interactive, conversational explanations?

The specific research questions addressed in the different papers in this dissertation are discussed in the section. 1.7.

1.4 Solution Approach

We propose the design of a user interface to provide personalized explanations in review-based RS, so users can explore explanatory information in an interactive manner, aiming to enhance user experience in terms of explanatory objectives, such as transparency, efficiency and trust in the system.

The design process described in this dissertation follows the principles of user-centered design, as defined by Vredenburg et al. (2002): "understanding of user requirements, iterative design and evaluation, and a multi-disciplinary approach". Our research methodology is summarized in 1.5.

Furthermore, and aiming to generate a theory-driven explanatory interface design, we set out to analyze explanations from the perspective of argumentation theory, which has produced a wide range of models of argumentation (Bentahar et al., 2010). This perspective can contribute to generating explanations that are properly grounded and structured, which can increase the likelihood of their understanding and acceptance.

One class of argumentation models defines - with many variations - logical structures of argumentative elements such as claims, evidence or facts supporting a claim, rebuttals and other components. A recommendation issued by a RS can be considered a specific form of a claim, namely that the user will find the recommended item useful or pleasing (Donkers & Ziegler, 2020). The role of an explanation is thus to provide supportive evidence (or rebuttals) for this claim. Claims are, however, also present in the individual user's rating and opinions, which may require explaining their grounds as well, thus creating a complex multi-level argumentative structure in an explainable RS. A different branch of argumentation theories (Walton & Krabbe, 1995) have abandoned the idea of static structural argumentation models and propose a dialectical approach to argumentation, focusing more on the process of exchanging arguments and supportive (or contradicting) information as part of a dialog between two parties, taking into account the social aspect of the explanatory process (an *explainer* transfers knowledge to an *explainee* (Miller, 2018)).

Motivated by the idea of explanations as a dialog between system and user, we propose an approach to explaining recommendations that allows users to interactively explore explanatory information at different levels of detail. Our approach to explanations as interactive argumentation involves an iterative process of 1) argumentation attempts: the system intends to provide arguments to explain recommendations, involving components such as claim, premise, backing, etc.; followed by 2) argument requests: the user asks the system to provide - follow-up - arguments that support the claim that user will find the recommended item useful (see Fig. 1.1).

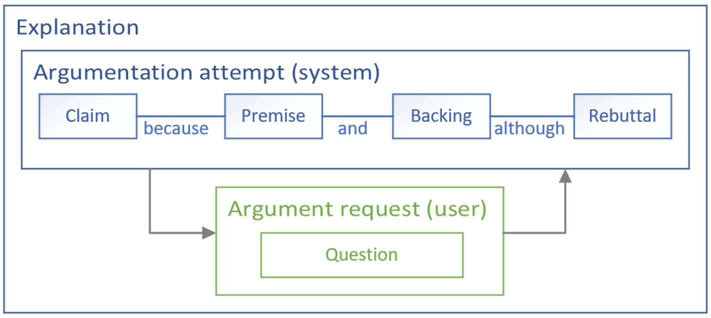


Figure 1.1: Overall scheme for explanations as interactive argumentation in review-based RS, as reported in Paper 6. Argumentation attempts are constituted by different argument components (claim, premise, etc.).

According to our proposal, the system takes the initiative and starts the dialog providing a premise of the type: "Great staff, very good location", which is included in the initial view of recommended options. With the above, we intend to offer an overall reason to support the implicit system claim of the system that the user will find the recommended item pleasant, and to make it easier for the user to select options - and reasons for recommendations - they want to examine in detail, taking into account that aspects of relevance to the user are highlighted in this first dialog move.

Our approach involve interactive options, that allow users to indicate when additional evidence is still needed to comprehend the system's explanatory claims. With respect to reviews, our approach also allows users to navigate from aggregated accounts of customer's opinions to detailed excerpts of individual reviews. Thus, our approach seeks to contribute to users' understanding of how the system works, and consequently to transparency and the general satisfaction with the system.

Consequently, we formulated the design of different types of explanatory interfaces to convey personalized explanations in review-based RS, inspired by argumentation models, both static and dialog-based, and the possibilities enabled by the method proposed by Zhang et al. (2014), EFM (Explicit Factor Model). This method requires the use of procedures to automatically detect aspect-based sentiments expressed by customers in their reviews, e.g. "I think the hotel staff was great" (aspect: *staff*, sentiment: positive).

To this aim, we relied on text classifiers using the state-of-art language model BERT (Devlin et al., 2019). By using the EFM, personalized explanations of the type: "We recommend this hotel because 95% of customers have made positive comments about the staff, an aspect that is relevant for you" can be provided. Here, personalization is reflected by a statement about the alignment of the user's preferences and the recommended item, in terms of aspects that may be relevant to the user. Thus, our proposal leverages the aspect-based focus of the EFM method to structure personalized explanations, in order to achieve a balance between explanations that are sufficient in content, but also brief and relevant to users.

To evaluate our approach, we first addressed the effect of providing explanations through different presentation styles (only text, table, bar chart) and involving different types of explanatory information (in regard to user preferences, item quality, and information on decision process), in a static fashion. We then addressed how interactivity could be leveraged to increase the positive users' evaluation of RS, in regard to explanatory aims, for which we designed and implemented both GUI navigation and natural language conversation interfaces, to allow users to explore explanatory information at will: in the former, by means of links and buttons; in the latter, by formulating explanatory questions in their own words, for which the conversational agent *ConvEx* was developed.

ConvEx falls into the finite-state dialog system approach, where the dialog state has a fixed set of possible transitions to other states. ConvEx includes a natural language understanding module, in charge of detecting intent expressed in user queries, trained on our own consolidated dataset ConvEx-DS, and leveraging BERT language model (Devlin et al., 2019).

To test the effects of our proposal, we performed a series of experiments taking as example the domain of hotels, since it represents an interesting mix between search goods (those with features on which complete information can be found before purchase or consumption (Nelson, 1981)) and experience goods (which cannot be fully known until purchase or consumption (Nelson, 1981)). Such a product evaluation could benefit from third-party opinions (Klein, 1998; Nelson, 1981), potentially rich in argumentative information that can be used for explanatory purposes.

1.5 Research Methodology

To achieve the overall aim of this dissertation, a user-centered design approach was followed, in which it was evaluated how users interact with different types of explanatory interfaces, as well as their opinion on the helpfulness of different interface components, explanation quality, transparency, effectiveness and trust. This research involved an iterative process, consisting on the following steps:

- Explanations and interface design: We formulated and refined a series of explanation schemes, based on argument theories and dialog models of explanation, and translated them into a subsequent interface design.

- Prototyping and development: To validate our design proposal with users, we developed mock-ups in early iterations of our research. Subsequent phases of our project involved the development of a complete explainable RS, and of a conversational agent for explainable RS.

- User-centric evaluation: We assessed users' evaluation of our proposal with online studies. To this end, we leveraged crowd-sourcing platforms (Amazon Mechanical Turk and Prolific) to test the effect of different presentation styles (such as text and graph-based), interface components (on users' profile and items' quality inference), and interface types (GUI navigation and natural language conversation).

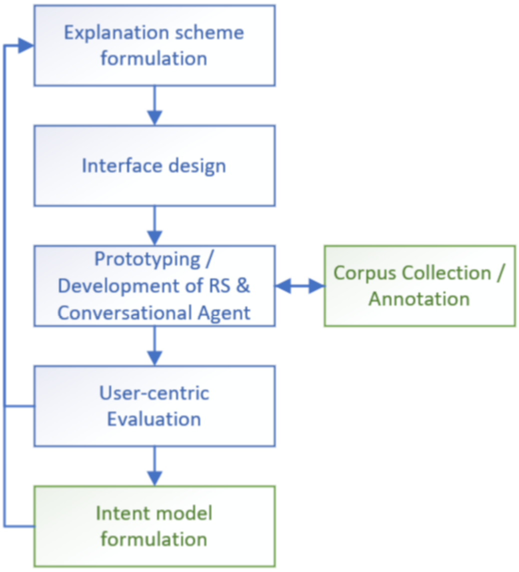

Explanation scheme formulation
Interface design
Prototyping / Development of RS & Conversational Agent
Corpus Collection / Annotation
User-centric Evaluation
Intent model formulation

Figure 1.2: Overview of research methodology. Blue boxes depict the main methodology flow (across all dissertation studies), while green boxes represent steps undertaken only for explanations as natural language conversation.

These steps are depicted in 1.2. Additionally, the development of a conversational agent to provide explanations through a natural language conversation interface involved the specific steps:

- User pre-study: As part of the conversational agent development, we performed an exploratory Wizard of Oz study, to gain insights into the type of questions users might ask in a conversational explanatory context.

- Formulation of an intent model: We modeled the need for information expressed in a user question (user intent) within a conversation with explanatory aim. This model served as a basis to refine our explanation scheme, and to implement procedures to automatically detect the intent of an explanatory query within a conversation.
- Corpus collection and annotation: We consolidated the ConvEx-DS, a dataset for the training of natural language understanding (NLU) procedures for automatic intent detection, with labels consistent to our intent model.

1.6 Evaluation Metrics and Measurements

This dissertation provides empirical evidence of the effect of our approach on users' evaluation of RS. In particular, we assessed metrics related to the explanatory aims defined by Tintarev (2007): transparency ("explain how the system works"), effectiveness ("help users make good decisions") and trustworthiness ("confidence in the system"). Thus, we assess users' subjective evaluation of RS in terms of system transparency, system effectiveness, trust in the system and explanation quality, to which we will refer as *transparency*, *effectiveness*, *trust* and *explanation quality* across this dissertation.

We utilized items from Knijnenburg et al. (2012) to measure effectiveness (system is useful and helps the user to make better choices), items from McKnight et al. (2002) to measure trust (constructs *trusting beliefs*, user considers the system to be honest and trusts its recommendations; and *trusting intentions*, user willing to share information to the system). We used the user experience items (UXP) of Kouki et al. (2019) to measure explanation quality, comprising items to address specific aspects of explanations, such as confidence, transparency of the explanation, and persuasiveness. We adapted an additional item from this scale, to measure satisfaction with the explanation. We also added an item adapted from Donkers et al. (2020) to evaluate explanation sufficiency.

In early iterations of our research, we used the item by Pu et al. (2011) to measure transparency (construct *transparency*, "I understood why the items were recommended to me"). However, in work we reported in Hellmann et al. (2022), we developed and validated a new instrument to measure user's perception of transparency, a questionnaire which was used during our latest user study (for which factor loadings and internal reliability were checked), aiming to report a most robust metric for transparency in contrast to the single item proposed by Pu et al. (2011), i.e. a set of questions involving further aspects related to RS transparency, such as input (e.g. understanding of data used by RS), functionality (e.g. system providing info on how and why is an item recommended), output (e.g. info on how well recommended items fit users' preferences), and interaction (e.g. understanding on what needs to be changed to get new recommendations). We note that our quantitative assessment focus on transparency as subjectively perceived by users (i.e. to what extent users believe they understand), rather than on objective transparency (i.e. the actual understanding of system's inner work by users (Rozenblit & Keil, 2002)). However, we also qualitatively address in Paper 6 the objective understanding of reasons behind recommendations, by analyzing participants responses to the open-ended question: "How would you explain to a friend, in your own words, how the system generates recommendations?".

As for user characteristics, we used the questionnaire proposed by Hamilton et al. (2016), which is a well-validated instrument for rational and intuitive decision making. Furthermore, we used the scale of the social awareness competency proposed by CASEL (2013), and the visualization familiarity items proposed by Kouki et al. (2019).

In regard to our consolidated ConvEx-DS, we calculated inter-annotator agreement (degree of agreement among independent human annotators), and accuracy of NLU procedures trained on the dataset, particularly F1 (the harmonic mean of the precision and recall), widely used to validate performance in supervised classification approaches, as is our case of intent detection procedures.

1.7 Specific Research Questions

This dissertation investigates how to provide explanations in review-based RS, so that the explanatory aims such as transparency, effectiveness and trust are achieved. Throughout this work, we explored how to formulate explanation schemes inspired by argumentation theory, as well as how to leverage more interactive explanations to increase the positive users' evaluation of RS. Particularly, we addressed the effect of different factors of explanation, namely: type of textual explanation based on reviews (aggregated results, only textual summary, using a helpful review), level of justification (high and low), visualization or display styles (text, table, bar chart), interface components (information on user preferences inference, item quality, information on decision process), the degree of interactivity to access explanatory information (high and low), and type of interface (GUI navigation and natural language conversation). Across this dissertation, users' evaluation is addressed in terms of the quality of explanations, and of the explanatory aims: transparency, effectiveness and trust, as defined by Tintarev (2007).Thus, this dissertation aims to answer the following specific questions:

RQ1 How do the type of textual explanation and the level of justification influence users' evaluation of RS? (Addressed in Paper 1).

RQ2 How do the presentation of different explanation interface components and different display styles influence users' evaluation of RS? (Paper 2 and Paper 3).

RQ3 How does access to explanatory information through different degrees of interactivity influence users' evaluation of RS? (Paper 3 and Paper 6).

RQ4 How do users communicate their explanation needs using a conversational agent? (Paper 4).

RQ5 How valid is our dimension-based intent model for explanatory user queries? (Paper 5).

RQ6 How does the use of different types of interactive explanatory interfaces influence users' evaluation of RS? (Paper 6).

RQ7 How do individual differences in user characteristics moderate users' evaluation of RS? (Papers 1, 2, 3 and 6)

Figure 1.3 summarizes the main concepts addressed in the research question of this dissertation. Each of the questions, as well as the results, are briefly presented below.

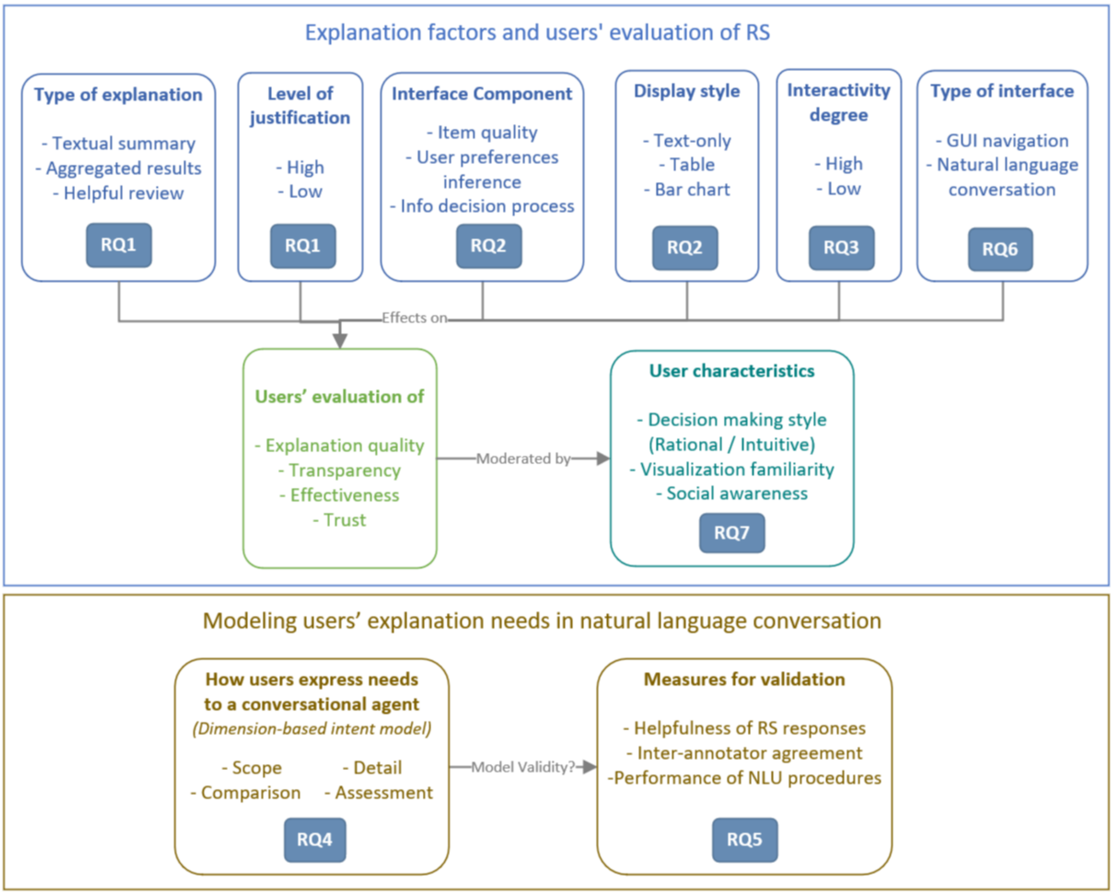


Figure 1.3: Overview of concepts addressed in specific research questions.

1.7.1 RQ1: How do the type of textual explanation and the level of justification influence users' evaluation of RS?

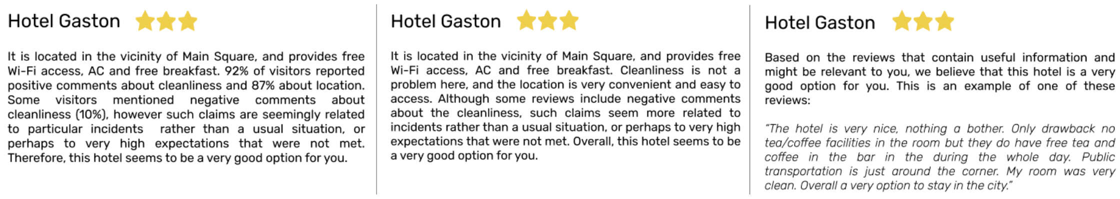
Problem: Fueled by advances in NLP, the interest in exploiting customer generated content has increased recently, not only to improve the accuracy of RS algorithms, but also to provide RS explanations by means of detailed opinions and sentiments towards different aspects of the item, instead of showing only general item properties, or numerical ratings granted by customers with similar preferences, as in most content-based or collaborative filtering explanation approaches. While a significant amount of work on review-based explanations is limited to providing a brief assessment of relevant aspects (e.g. "You might be interested in [feature], on which this product performs well" (Zhang et al., 2014)) or aggregated views of pros and cons (e.g. explanations using a bar chart visualization as in Muhammad et al. (2016)), we hypothesized that users in need of further information may be more satisfied when more detailed arguments or a higher level of justification for the recommendations is provided.

However, and due to a general lack of user-centered evaluation of RS explanatory approaches and methods (Nunes & Jannach, 2017), we encountered a number of unresolved questions related to review-based explanations: Do users prefer concise explanations over those that include more specific details? Do they prefer an aggregated view of other users' opinions, over reading individual reviews written by similar users? We hypothesized, for example, that users would rate better explanations as an aggregated view of opinions (e.g. with percentages of positive/negative opinions), than those involving a textual summary, as numerical figures could function as anchors that might convey more compelling information about the items.

While a prevalent approach to evaluate the quality of explanations generated in natural language by a system is the use of offline evaluation metrics (e.g. BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004)), it became necessary to empirically evaluate - with users - the quality of explanations generated under different review-based explanatory methods, which could point us in the proper direction as to the explanatory method to base our further developments on. Thus, we first investigated the effect of review-based explanations under different types of textual explanation (aggregated opinions, only textual summary, using a helpful review), and different levels of justification (high and low). Paper 1 addresses this question, following the methodological steps described in section 1.5.

Approach: To answer this question, we first formulated an explanation design with an argumentative structure, inspired on the scheme proposed by Habernal & Gurevych (2017), a variation of original Toulmin's model (Toulmin, 1958), that seeks to represent the kind of arguments usually provided in user-generated web discourse. This first explanation involves the argument components: premise (item attributes), claim (hotel seems to be a good option for the user), backing (report of positive opinions), rebuttal (report of negative opinions) and refutation (possible reason for negatives). Based on this scheme, a series of templates reflecting the following three types of review-based explanation where created (see examples of explanations in Fig. 1.4):

- Explanations with aggregated results: An accumulated view using bullet points and percentages of positive and negative opinions per aspect, as proposed by Gerani et al. (2014).
- Explanations with only textual summary: Summarization of opinions without bullet points nor percentages. It resembles a system generated review, as proposed by F. Costa et al. (2018), and Carenini et al. (2013).
- Explanations using a helpful review: Indicate that the recommendation was based on the reviews that might be helpful to the user, as proposed by C. Chen et al. (2018), and show just one of them as an example.



Hotel Gaston ★★★

It is located in the vicinity of Main Square, and provides free Wi-Fi access, AC and free breakfast. 92% of visitors reported positive comments about cleanliness and 87% about location. Some visitors mentioned negative comments about cleanliness (10%), however such claims are seemingly related to particular incidents rather than a usual situation, or perhaps to very high expectations that were not met. Therefore, this hotel seems to be a very good option for you.

Hotel Gaston ★★★

It is located in the vicinity of Main Square, and provides free Wi-Fi access, AC and free breakfast. Cleanliness is not a problem here, and the location is very convenient and easy to access. Although some reviews include negative comments about the cleanliness, such claims seem more related to incidents rather than a usual situation, or perhaps to very high expectations that were not met. Overall, this hotel seems to be a very good option for you.

Hotel Gaston ★★★

Based on the reviews that contain useful information and might be relevant to you, we believe that this hotel is a very good option for you. This is an example of one of these reviews:

"The hotel is very nice, nothing a bother. Only drawback no tea/coffee facilities in the room but they do have free tea and coffee in the bar in the during the whole day. Public transportation is just around the corner. My room was very clean. Overall a very option to stay in the city."

Figure 1.4: Example of explanations provided in user study reported in Paper 1, under the condition low level of justification. Left: explanation type aggregation; middle: summary; right: helpful review.

Additionally, per each type of explanation, templates reflecting the following levels of justification were created:

- High: To address the main aspects of interest to users (e.g. overall cleanliness) by providing fine grained details with several sentences about more specific aspects (e.g. cleanliness of bathroom or carpet).
- Low: To address the main aspects of interest to users (e.g. overall cleanliness), without further elaboration or details, see example Fig. 1.4.

Based on the above mentioned templates, we generated a mock-up prototype, which allowed users to navigate through a list of fixed recommendations, and its explanations. We then conducted our first user study, based on a between subjects factorial design (3x2), where every participant was randomly assigned to a condition representing the combination of the factors: type of explanation and level of justification, as described in detail in Paper 1. Users' evaluation was measured in terms of explanation quality and explanatory aims (transparency, trust, etc.)

Outcomes: We found the type of textual explanation influences significantly the perception of explanation quality. In particular, explanations that provided percentages of positive and negative opinions were reported as significantly better than textual summaries without any percentages. Second, we found that the type of explanation significantly influences the evaluation of transparency. In particular, we found that the prototype was evaluated significantly as more transparent when it provided explanations as aggregated opinions, than when only a helpful review was provided. Consequently, we decided to focus on methods that facilitate calculation of percentages of positive and negative opinions, which can in turn be supported by examples (extracted from reviews) that constitute such aggregated statistics. We then based our further developments on the EFM, the explanatory RS method proposed by Zhang et al. (2014), which facilitates the generation of aspect-based statistics of opinions.

In particular, we explore in Paper 2 and Paper 3 possibilities and effects of graphical visualization of different types of interface explanatory components, such as user preferences, item quality and information on decision process. Furthermore, we based the studies reported in Papers 3, 4 and 5 on our implementation of the EFM method, to further test differences in display styles, types of interface (GUI navigation and natural language conversation) and degrees of interactivity, that allowed participants to explore arguments involving also excerpts from users reviews in connection to aggregated opinions.

Contrary to our expectations, we found no main effects of the level of justification. At this point, we noted that providing more detailed explanations through a single static presentation of arguments does not necessarily imply a more positive evaluation of the system and its explanations. Here, even detailed explanations may leave out momentary relevant aspects to users, or aspects that raise attention due to many reported negative comments. Without access to this additional information, doubts about the item suitability may persist. And since providing details on all aspects in a single view can be both overwhelming and unnecessary, a more flexible solution is needed to explore aspects and levels of detail at will, a topic explored in Paper 3 and Paper 6, where interactive mechanisms for such desiderata are addressed.

1.7.2 RQ2: How do the presentation of different explanation interface components and different display styles influence users' evaluation of RS?

Problem: Review-based explanations based on aggregated statistics focus mainly on an item's quality view, i.e. aspect-based pros and cons, neglecting mostly the details on how the user profile is calculated. Our chosen explanatory method, EFM by Zhang et al. (2014), infers user preferences based on the frequency on which the user has addressed aspects on their own reported reviews, in order to predict item ratings according to such preferences. While providing a view on the alignment between preferences and item features is common in content-based approaches (as in He et al. (2015); Vig et al. (2009)), it was not clear how this alignment should be presented in review-based approaches so details on user's profile inference are also shown, nor to what extent providing such information might be considered useful by users, in comparison with other explanation components as item quality (pros and cons), and information on the decision process. In this regard, we hypothesized that users would benefit from an explanatory view reporting on details on how their preferences were calculated.

Aggregated opinions on positive and negative aspects could be displayed using different styles. A very popular approach is to display pros and cons using bar charts (e.g. Muhammad et al. (2016)), which may be regarded as more informative and appealing than brief textual explanations, easier to interpret than challenging visualizations as radar charts (e.g. Wu & Ester (2015)), or faster to process than tabulated data. However, it was not clear to what extent the presentation format could influence the evaluation of review-based RS and their explanations. We hypothesized, for example, that explanations providing a summarized view on customer opinions through a bar chart would be preferred, particularly by users more familiar with graphic representations of information, given to their greater possibility of quick processing.

Approach: We based on the explanation scheme introduced in Paper 1, and proposed, in Paper 2, an explanation design consisting of propositions backing (and rebutting) the claim that an item is worth being chosen. We proposed the following explanation interface components (see example in Fig. 1.5):

- Information on item quality: A summary of comments reported by previous hotel guests for different aspects that may be relevant to users, as well as what percentage were positive and negative.
- Information on user preferences inference: explicit data regarding user profile inference, particularly, the number of comments reported by the user about an aspect in their previous comments.
- Information on decision process: Statements that inform how the user preferences and item quality are extracted (e.g,"based on how often you mentioned these features in your own comments before").

By including the user preference component we aimed to make the user's own profile transparent, by showing the user's inferred importance of each aspect (to what extent the user addressed an aspect on their own previous reviews), together with reported customer comments on those aspects, so a direct comparison of the points of view of others and their alignment with their own preferences was facilitated. By providing information on the decision process component, we aimed to improve users' evaluation

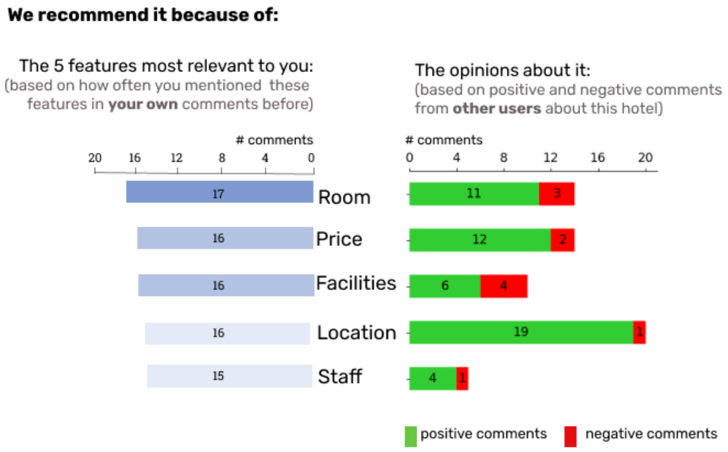


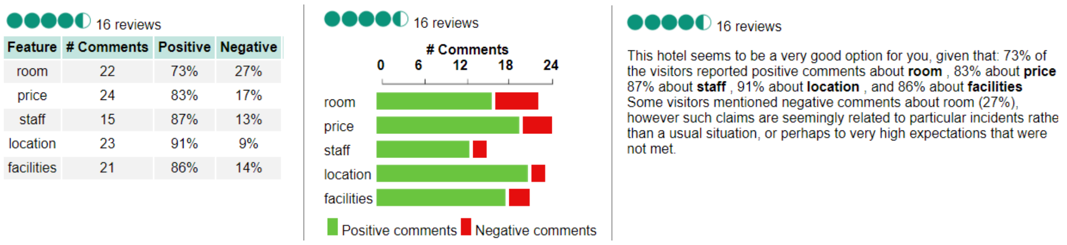
Figure 1.5: Example of explanation provided in user study reported in Paper 2, involving the interface components: information on item quality (bar chart right), on user preferences inference (bar chart left) and on decision process (text: "based on...").

in regard to explanatory aims, by indicating explicitly how the item quality and the users' own profile were calculated.

While arguments are usually associated with oral or written speech, arguments can also be communicated using visual representations, such as graphics or images. According to Blair (2012), visual arguments (a combination of visual and verbal communication) may, in addition to representing propositional content, have a greater "rhetorical power potential" than verbal arguments, due to their greater immediacy, i.e. possibility of quick processing.

In consequence, we described in Paper 2 two experiments, with which we aimed to test the effect of the two factors: display style (using a bar chart, or a table) and the display of detailed information on the inference of user preferences (yes, no). In the first experiment, we used a between-subjects factorial design, where each participant was assigned to a condition reflecting the combination of the 2 factors (display style and display of the inference of user preferences). To this aim, we generated a mock-up prototype, which allowed users to navigate through a list of fixed recommendations, and its explanations. Users' evaluation was measured in terms of explanation quality, and explanatory aims (transparency, trust, etc.). In a second experiment, we used a within-subjects design, in which each participant was exposed to the 4 types of explanations (combination of the two factors, display style and display user preferences), with the aim of observing possible differences in evaluation at individual level, and to examine further into specific aspects of explanations (e.g. explanations' ease of understanding, or explanations' persuasiveness). In addition, this experiment also involved the assessment of the evaluation of helpfulness of individual components of explanations, by users.

We investigated further in Paper 3 the impact of different presentation styles, including, in addition to table and bar chart styles, the comparison with text-only explanations (see Fig. 1.6).



16 reviews

Feature	# Comments	Positive	Negative
room	22	73%	27%
price	24	83%	17%
staff	15	87%	13%
location	23	91%	9%
facilities	21	86%	14%

16 reviews

This hotel seems to be a very good option for you, given that: 73% of the visitors reported positive comments about **room** , 83% about **price** 87% about **staff** , 91% about **location** , and 86% about **facilities** Some visitors mentioned negative comments about room (27%), however such claims are seemingly related to particular incidents rathe than a usual situation, or perhaps to very high expectations that were not met.

Figure 1.6: Example of explanations provided in user study reported in Paper 3, involving the presentation styles: table (left), bar chart (middle), text-only (right).

Outcome: Contrary to our expectation, we found no main effect of the display of details about the inference of user preferences, on users' evaluation of the system or its explanations. However, when seeking further into specific aspects of explanations, we found a multivariate significant effect of display of the inference of user preferences. Particularly, we found that explanations that do not include detailed information on the inference of user preferences were significantly easier to understand, compared to those including such a component. Additionally, while most participants reported they found the item quality component useful, the opposite was the case for the component about own preferences' inference, with only a minority of participants reporting they found this section useful. The above suggests that, at least for the domain we chose as example (hotels), users seemed to be much more interested in other people's opinion and their weight in the recommendation, rather than how their own profile was calculated. Consequently, in further experiments reported in Paper 3 and Paper 6, we omitted the interface component on detailed information about user profile inference.

In regard to effects of display style, we did not find a main effect of this factor on the evaluation of the explanations or the system by users, unless users' characteristics are taken into account, as we will discuss in RQ7.

1.7.3 RQ3: How does access to explanatory information through different degrees of interactivity influence users' evaluation of RS?

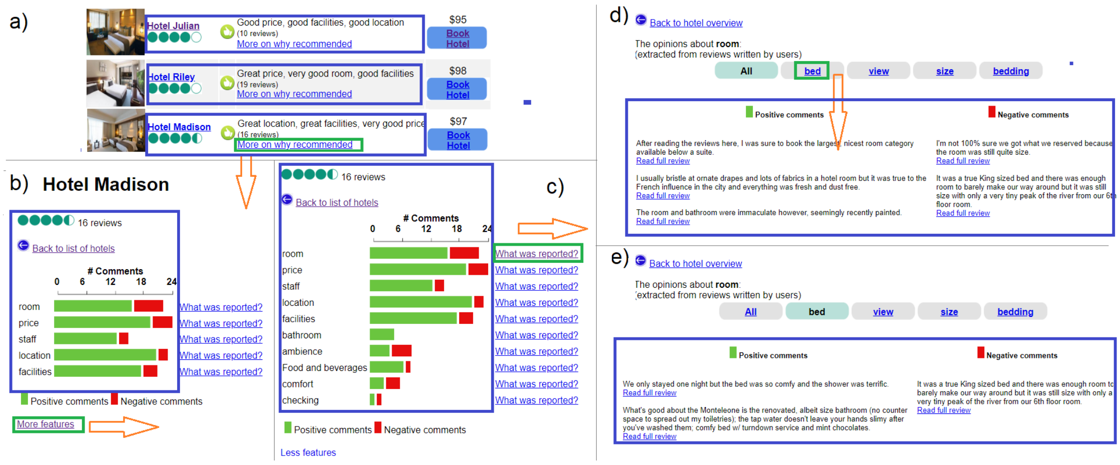
Problem: The predominant explainable RS involving customer opinions is to provide explanations in a static manner (i.e., using a non-interactive presentation), limiting users in exploring the diverse views and arguments expressed in the reviews, which in turn can support system attempts at explanation. While providing interactive explanations may positively impact users' evaluation of RS, explanatory methods that allow users to scrutinize and customize explanations through interaction are largely unexplored, or lack sufficient empirical evidence (Sokol & Flach, 2020). In particular, in the context of review-based RS, one question that remained open was the extent to which users' evaluation vary when different degrees of interactivity are enabled to access explanatory information at various levels of detail. In this regard, we hypothesized that users' evaluation of the RS and its explanations is more positive when explanations are provided with a higher degree of interactivity.

Approach: To answer this question, we followed the research methodology steps described in 1.5, and reported in Paper 3 and Paper 6. First, we formulated an explanation scheme, reflecting the concept of explanations as interactive argumentation, grounding on argumentation theory and dialog models of explanation, seeking to facilitate the exploration of arguments that support the claims made by the system, while providing answers to their explanation-related questions, at different levels of detail. Here, we regard an explanation as an iterative process, i.e. as a sequence of argumentation attempts (the system intends to provide arguments to explain something) followed by argument requests (the user asks the system to provide - follow-up - arguments that support the claim that the user will find the recommended item useful. In turn, the argumentation attempt to requests of the type "why-recommended" reflects, on a first level, the argument structure discussed in Paper 2, to provide an aggregated view of pros and cons per aspect.

Based on the explanation scheme, we formulated the design of two interface types, namely GUI navigation (Paper 3, see example in Fig. 1.7), and natural language conversation (Paper 6, see example in Fig 1.8). Here, design features such as links and buttons (GUI navigation case) and user written questions and buttons (natural language conversation case) foster the following interactive features (as defined by Y. Liu & Shrum (2002)): active control, by enabling users to be in control of which argumentative content to display; and two-way communication by enabling users to indicate the system which argumentative statements require further elaboration, and which features are of real relevance at the time of making the decision.

To answer this RQ, and unlike the studies in Paper 1 and Paper 2 (where mock-up prototypes were used), we developed a base RS based on the EFM method (Zhang et al., 2014), and sentiment-based aspect detection procedures, using the state of art natural language processing model BERT (Devlin et al. (2019)), to generate matrices of user preferences and items quality. In the first instance, we developed an GUI navigation interface (Paper 3), and conducted a user study to compare users' evaluation of the system, in terms of explanatory aims (transparency, effectiveness and trust), and of the specific aspects of explanations (explanation confidence, persuasiveness, sufficiency, etc.). We examined the effect of the degree of interactivity: high (further options to access detailed arguments are provided) and low (far fewer options to explore explanatory content).

Next, we extended the implemented RS, and further integrated it to a conversational agent (Paper 6), with a view to test the effect of different degrees of interactivity in regard to explanations as natural language conversations, and also to compare with the GUI navigation alternative (see RQ6).



Hotel Julian: Good price, good facilities, good location (10 reviews) More on why recommended

Hotel Riley: Great price, very good room, good facilities (19 reviews) More on why recommended

Hotel Madison: Great location, great facilities, very good price (16 reviews) More on why recommended

d) The opinions about **room**:
(extracted from reviews written by users)

🟩 Positive comments

After reading the reviews here, I was sure to book the largest, nicest room category available below a suite.
Read full review

I usually bristle at ornate drapes and lots of fabrics in a hotel room but it was true to the French influence in the city and everything was fresh and dust free.
Read full review

The room and bathroom were immaculate however, seemingly recently painted.
Read full review

🟥 Negative comments

I'm not 100% sure we got what we reserved because the room was still quite size.
Read full review

It was a true King sized bed and there was enough room to barely make our way around but it was still size with only a very tiny peak of the river from our 6th floor room.
Read full review

e) The opinions about **room**:
(extracted from reviews written by users)

🟩 Positive comments

We only stayed one night but the bed was so comfy and the shower was terrific.
Read full review

What's good about the Monteleone is the renovated, albeit size bathroom (no counter space to spread out my toiletries); the tap water doesn't leave your hands slimy after you've washed them; comfy bed w/ turndown service and mint chocolates.
Read full review

🟥 Negative comments

It was a true King sized bed and there was enough room to barely make our way around but it was still size with only a very tiny peak of the river from our 6th floor room.
Read full review

Figure 1.7: Interactive explanations through GUI navigation (screenshots of implemented system, reported in Paper 3, user study condition high interactivity. Enclosed in blue: argumentation attempts; in green: argument requests. Orange arrows: sequence of allowed moves, pointing to the next interface.

Outcome: Our results show that a higher degree of interactivity has a significantly positive effect on users' evaluation, in terms of explanation quality and explanatory aims, compared to explanations with a lower degree of interactivity, in both GUI navigation and natural language conversation interfaces. The above in turn confirms the suitability of our proposal for explanations as interactive argumentation, inspired by conversational models of explanation, which enables users to contest initial explanation attempts provided by the system, in particular the aggregate representation of positive and negative customer opinions.

1.7.4 RQ4: How do users communicate their explanation needs using a conversational agent?

Problem: The type of interaction most prevalent in interactive approaches in RS and, more broadly, in explainable artificial intelligence, is the GUI navigation style, where interaction is enabled by point-and-click options. However, recent advances in NLP allow for further interaction possibilities, in which users could indicate, in their own words, their needs for explanation, as part of a conversation, instead of being limited to fixed questions pre-determined by the system.

Although conversational interfaces have been developed and used for a long time to assist users in a wide range of tasks, little is known about how a conversational agent should be designed in the context of explainable review-based RS. In particular, little is known about the type of explanation-related questions users would ask to a RS. Thus, in order to develop our conversational agent for explainable RS, it was first necessary to understand how users would formulate their explanatory queries to a conversational agent.

Approach: To answer this question, we conducted a pre-study, using the Wizard of Oz (WoOz) methodology (Kelley, 1984), as reported in Paper 4, that provided insights into the type of questions users might ask to a conversational agent. Such technique allows to validate how users would interact with a conversational interface, and to evaluate the feasibility of dialog based systems that have not yet been fully implemented. Under this paradigm, a human-machine interaction is simulated, in which a member of the research team (the wizard) executes the response actions on behalf of the system, through a computer-mediated interface, which we developed to this aim.

Outcome: Based on our observations, we formulated a dimension - based intent model. We identified that users' intents could be classified into the types: *domain-related* intents (regarding hotels and their features), and *system-related* intents (regarding the algorithm, the system input, or system functionalities). In turn, domain-related intents could be categorized according to the following dimensions:

- Scope: Whether the question refers to a single item (*single*), a limited list of items (*tuple*), or to no particular item (*indefinite*).
- Comparison: Whether the question is (*comparative*) or not (*non-comparative*).
- Assessment: Whether the question refers to the existence or characteristics of item features (*factoid*), to a subjective assessment of the item or its features (*evaluation*), or to system reasons to recommend an item (*why-recommended*).

- Detail: Whether the question inquires for a specific aspect or feature (*aspect*), or for the overall item (*overall*).

Consequently, the intent of a single domain question can be defined as a combination of the four dimensions. This dimension-based intent model served as basis for the corpus annotation guidelines (for collection of ConvEx-DS and the addressing of RQ5), as well as for the design of the dialog policy, to build our conversational agent for explainable RS (to address RQs 3, 6 and 7); in this latter case, possible conversation flows were defined, according to the intent recognized for a user question.

1.7.5 RQ5: How valid is our dimension-based intent model for explanatory user queries?

Problem: To our knowledge, there were no publicly available datasets intended to support the development of an explanatory conversational agent for RS, nor datasets for detecting user intent expressed in a question in such a context, as discussed in detail in Paper 4 and Paper 5. As mentioned, the evaluation of RQ4 resulted in the formulation of an intent model for this purpose. However, by means of the WoOz study, only a low number of users' questions was obtained, so it was necessary to evaluate the validity of the model to a larger scale, i.e. the extent to which the model is able to accurately represent user intents.

Approach: As an indirect measure of validity, we set out to evaluate helpfulness of the responses generated by a RS implementing the intent model, under the assumption that if the system has adequately recognized the user's intent, it is able to generate a response that approximates the user's information need, and thus be considered, to some extent, helpful. To further evaluate the validity of the model, we aimed to test to what extent the collected questions could be consistently classified by human annotators.

To this end, we extended the base RS we implemented to answer RQ3, by implementing a module to interpret user queries in natural language, and to provide answers based on the underlying RS algorithm used (EFM, Zhang et al. (2014)). Here, we relied on text classifiers for the different intent dimensions, using the state-of-the-art language model BERT (Devlin et al., 2019) and auxiliary datasets that were useful for detecting certain (but not all) dimension values. We then conducted a user study aiming both to collect a large number of user queries, and to measure the helpfulness of system generated answers to users.

Next, the intent of the collected questions was annotated, using guidelines inspired by the intent model definition (outcome of RQ4). We consolidated the intent gold standard for each question, and validated the performance of intent detection procedures trained using the final annotated corpus.

Outcome: We published ConvEx-DS, a dataset with 1806 questions [2], containing annotations to train intent detection procedures, that facilitate the development of conversational agent for explainable review-based RS, basis to address our RQ3 and RQ6. In regard to the model validity, we found a substantial annotation agreement for each dimension, as well as a very encouraging accuracy of automatic classifiers, when these were trained on the ConvEx-DS. The above indicates that under our proposed intent model and annotation guidelines, users' questions could be, to a substantial extent, unequivocally classified, both by humans and by automatic classification procedures.

Furthermore, we found that the system was able to generate an answer in more than 80% of the cases, and to partially recognize the intent or entities in 7.34% of the cases (thus asking the user to rephrase or indicate further information). Finally, we found that system answers were regarded as predominantly helpful by users.

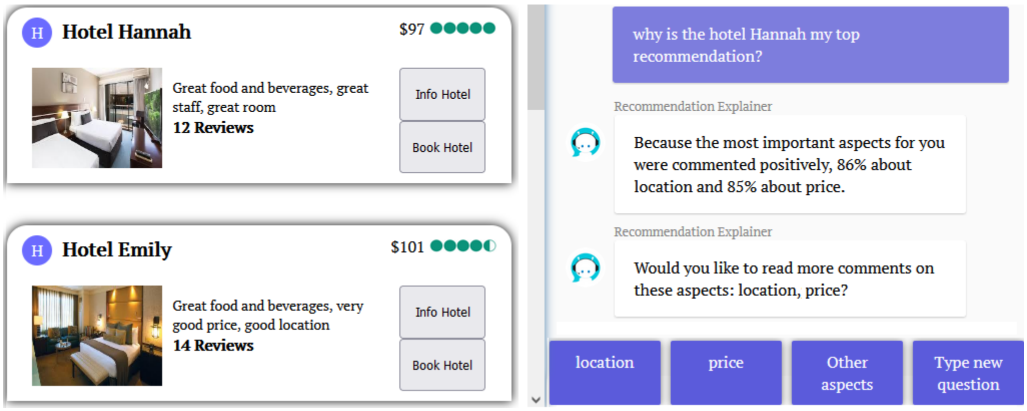
1.7.6 RQ6: How does the use of different types of interactive explanatory interfaces influence users' evaluation of RS?

Problem: Conversational user interfaces have been developed and used for a long time to assist users in a wide range of tasks (Moore & Arar, 2018), including applications for user preference elicitation in RS, as well as related processes, such as booking or purchasing. However, little is known about the advantages (and disadvantages) of a natural language conversation approach over an GUI navigation one, from the users' point of view, in the context of RS. In particular, and to our knowledge, there has been no empirical evaluation comparing the users' evaluation of RS providing explanations using such interactive approaches. In this regard, we hypothesized that users' evaluation of the system would be more positive when providing explanations through a natural language interface, than when provided through a GUI navigation, given that the former enables the user to formulate a wider range of questions, even in their own words, compared to the GUI navigation approach, where questions were limited to a much smaller set of query options.

Approach: Following the user-centered design approach described in Section 1.5, we implemented the conversational agent for explainable RS ConvEx (see Fig. 1.8). This system includes a natural language understanding (NLU) module trained on ConvEx-DS, and a dialog policy designed to enable conversation flows given different types of recognized intents, which in turn is based on an explanatory scheme resulting from previous iterations (when addressing RQs 1 to 4), and adapted to support possible flows of an explanatory conversation. Details of ConvEx implementation and evaluation are addressed in Paper 6.

[2] ConvEx-DS can be downloaded at https://github.com/intsys-ude/Datasets/tree/main/ConvEx-DS

ConvEx allows not only to answer standalone questions, but also to provide additional arguments to support the system's responses to user requests, through interaction options that facilitate users' input of follow-up questions, as well as triggering other system actions as a result of the conversation, e.g. highlighting options that include features the user is asking about. Subsequently, we conducted a user study to compare users' evaluation of both interactive interfaces for explainable RS (GUI navigation and natural language conversation), in terms of explanation quality and explanatory aims (transparency, effectiveness and trust).



H Hotel Hannah $97 ●●●●●

Great food and beverages, great staff, great room

12 Reviews

H Hotel Emily $101 ●●●●◐

Great food and beverages, very good price, good location

14 Reviews

why is the hotel Hannah my top recommendation?

Recommendation Explainer

Because the most important aspects for you were commented positively, 86% about location and 85% about price.

Recommendation Explainer

Would you like to read more comments on these aspects: location, price?

Figure 1.8: Interactive explanations through a natural language conversation (screenshot of implemented system ConvEx), reported in Paper 6.

Outcome: Providing explanations in RS using both types of interface (GUI navigation and natural language conversation) proved to be a meaningful means of benefiting user experience, given the predominant positive evaluation of the two tested systems and their explanations by the participants. However, we found no significant difference in the evaluation of the two types of interfaces, unless user characteristics are taken into account, which is addressed in our RQ7. Based on our findings, we discuss in Paper 6, a series of trade offs related to the use of one or the other type of interfaces, and we finish Paper 6 with a section of practical implications, aimed at designers in the field of RS.

1.7.7 RQ7: How do individual differences in user characteristics moderate users' evaluation of RS?

Problem: Individual user characteristics can influence the evaluation of a RS (Knijnenburg et al., 2012; Xiao & Benbasat, 2007). Additionally, and regardless of its type or disposition of components, an explanation may not satisfy all possible explainees

(Sokol & Flach, 2020). In consequence, and in line with Y. Liu & Shrum (2002)), we hypothesized that a number of user characteristics may moderate the effect of the different presentation and interactive features discussed in RQs 1, 2, 3 and 6, on the perception of explanation quality and the overall system. While previous authors addressed the effect of personality traits (e.g. the Big-Five traits in (P. T. Costa & McCrae, 1992; Tkalcic & Chen, 2015)), we aimed to evaluate how user characteristics differences related to decision-making styles, social awareness and visualization familiarity, moderate the effect of different types of explanation, display of the inference of user preferences, presentation styles, and types of interactive interfaces, on users' evaluation of review-based RS.

Approach: To address this question, we included questionnaires related to the above mentioned user characteristics, in user studies reported in Papers 1, 2, 3 and 6.

Outcome: We observed that differences in user characteristics may lead to differences in evaluation of our overall approach to explanations. Particularly, we observed main effects of factors such as rational decision-making style and social awareness, as well as interaction effects between: social awareness and type of explanation; social awareness and display of the inference of user preferences; intuitive decision-making style and display style; social awareness and degree of interactivity; and visualization familiarity and interface type.

1.8 Contributions and Related Publications

This section summarizes the main contributions of this dissertation, following the work reported in Papers 1 to 6. An overview of the contributions is depicted in table 1.1.

Schemes for explanations as static argumentation in review-based RS were defined, inspired by argumentation theory, in order to provide review-based arguments backing RS claims, under different display styles, as discussed in Paper 1 and Paper 2.

Schemes for explanations as interactive argumentation in review-based RS were formulated. Such design was inspired by argumentation theory and dialog based models, in order to provide review-based arguments that support RS attempts at explanation, and to allow users to challenge the system for further arguments when these are not yet fully understood or accepted. Varied interface designs and interaction flows involving GUI navigation and natural language conversation interfaces were designed based on these schemes, as reported in Paper 3 and Paper 6.

	Paper					
	1	2	3	4	5	6
Research Question						
RQ1 Effects of type of textual explanation and level of justification	•					
RQ2 Effects of explanation interface components and display styles		•	•			
RQ3 Effects of degree of interactivity			•			•
RQ4 How do users communicate their explanation needs to a conv. agent				•		•
RQ5 How valid is our dimension-based intent model					•	
RQ6 Effects of interactive interface type						•
RQ7 Effects of individual differences in user characteristics	•	•	•			•
Contributions						
Schemes for explanations as static argumentation	•	•				
Schemes for explanations as interactive argumentation			•			•
Dimension-based intent model				•		
Dataset for automatic intent detection					•	
Explainable review-based RS implementation			•			
Conversational agent implementation						•
Empirical evaluation (user studies)	•	•	•	•	•	•

Table 1.1: An overview of the contributions of this work, addressed in each paper.

A dimension-based intent model for users' questions in conversational explanations in review-based RS was formulated, which represents the type of questions that users might ask to a conversational agent with explanatory purposes in the hotel domain. This model served as a basis for the collection and annotation of a corpus involving such type of users' queries, and for the formulation of a dialog policy, required for the development of a conversational agent for explanatory review-based RS, as discussed in Paper 6.

A dataset for automatic intent detection in conversational explanations in review-based RS was released, including 1806 intent annotations for user questions with explanatory purpose in the domain of hotels, which can facilitate the development of explanatory conversational agents in RS, as addressed in Paper 5.

An explainable review-based RS was developed, based on the EFM algorithm, and aspect-based sentiment procedures using the state of art natural language processing model BERT (Devlin et al., 2019).

A conversational agent for explanatory review-based RS was developed, as addressed in Paper 6. The ConvEx system is able to answer different types of users questions, including comparative, factoid, and why-recommended questions. It enables conversation flows for different types of intent, and enables users to access additional arguments that support system explanation attempts, at will. ConvEx allows not only to answer standalone questions, but also provides interaction options to facilitate users' input of follow-up questions, as well as triggering other system actions as a result of the conversation, e.g. highlighting options including features relevant to the user.

Empirical evaluations We conducted a total of 7 user studies, in order to test our approach to static and interactive review-based explanations. Particularly, we tested: differences between types of textual explanation and levels of justification (n=152, Paper 1), differences between display styles and explanation interface components (n=150, n=35, Paper 2), differences between display styles and interactivity degrees (n=170, Paper 3), types of questions users would ask to an explanatory conversational agent (n=20, Paper 4), types of questions written by users and helpfulness of responses generated by conversational explanatory RS (n=298, Paper 5), and differences between different interactivity degrees and types of explanatory interfaces - GUI navigation and natural language conversation - (n=162, Paper 6).

2 Papers Contained in the Dissertation

2.1 Paper 1

This paper was published as:

Hernandez-Bocanegra, D.C., Donkers, T., & Ziegler, J. (2020). Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*, 219-225. ACM, New York, NY, USA. doi: https://doi.org/10.1145/3386392.3399302

This paper presents a user study conducted to test the effect of different types of textual explanation based on review-based explanatory methods. We compare in this paper users' evaluation of explanations providing aggregated opinions, a textual summary of opinions, or a helpful review. Also, we evaluate in this paper the effect of different levels of justification, namely high and low (more or less details on general / fine grained aspects). This paper addresses the research questions RQ1 and RQ7.

Contributions as first author: I was responsible for the design and generation of user interface mock-ups, user study setup and execution, data analysis and the writing of the paper. The design, methodologies and results of this study were discussed with co-authors, who also contributed extensively to the revision of the text.

2.2 Paper 2

This paper was published as:

Hernandez-Bocanegra, D.C., & Ziegler, J. (2020). Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics. *Journal of Interactive Media, i-com*, 19(3), 181–200. doi: https://doi.org/10.1515/icom-2020-0021

This paper presents two user studies conducted to test the effect of providing review-based explanations under different display styles (namely table and bar-chart), as well as the effect and helpfulness of different explanation components (namely details on the inference of user preferences, item quality, information on decision process). This paper addresses our questions RQ2 and RQ7.

Contributions as first author: I was responsible for the design and generation of user interface mock-ups, user study setup and execution, data analysis and the writing of the paper. The design, methodologies and results of this study were discussed with Prof. Dr. Jürgen Ziegler, who also contributed extensively to the revision of the text.

2.3 Paper 3

This paper was published as:

> Hernandez-Bocanegra, D.C., & Ziegler, J. (2021). Effects of Interactivity and Presentation on Review-Based Explanations for Recommendations. In: Ardito C. et al. (eds) Human-Computer Interaction – INTERACT 2021. INTERACT 2021. Lecture Notes in Computer Science, vol 12933. Springer, Cham. https://doi.org/10.1007/978-3-030-85616-8_35

This paper addresses the concept of explanations as interactive argumentation, and introduces the design of an interactive interface to provide explanations under the GUI navigation paradigm. This paper also presents a user study, where we tested the effect of providing review-based explanations under different display styles (namely text, table, bar-chart), as well as the effect of different degrees of interactivity to provide explanatory information (namely low and high). This chapter addresses our questions RQ2, RQ3 and RQ7.

Contributions as first author: I was responsible for the development of the explainable review-based RS, the aspect-based sentiment detection methods and the user interface. I was also responsible user study setup and execution, data analysis and the writing of the paper. The design, methodologies and results of this study were discussed with Prof. Dr. Jürgen Ziegler, who also contributed extensively to the revision of the text.

2.4 Paper 4

This paper was published as:

Hernandez-Bocanegra, D.C., & Ziegler, J. (2021). Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *3rd Conference on Conversational User Interfaces (CUI '21)*, 1-11. ACM, New York, NY, USA. doi: https://doi.org/10.1145/3469595.3469596

This paper presents a user pre-study, with which we aimed to obtain insight about how users would indicate their explanatory needs to a conversational agent, for which we relied on the Wizard of Oz paradigm. We introduce in this paper our proposal of a dimension-based intent model, which aims to facilitate the development of intent detection procedures for explainable RS in the domain of hotels. This paper addresses our question RQ4.

Contributions as first author: I was responsible for the development of system functionalities required for the user study. I was also responsible for the user study setup and execution, data analysis and the writing of the paper. The design, methodologies and results of this study were discussed with Prof. Dr. Jürgen Ziegler, who also contributed extensively to the revision of the text.

2.5 Paper 5

This paper was published as:

Hernandez-Bocanegra, D.C. & Ziegler, J. (2021). ConvEx-DS: A dataset for conversational explanations in recommender systems. In *Proceedings of IntRS 21: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 1-18. url: http://ceur-ws.org/Vol-2948/paper1.pdf

We report in this paper the collection and annotation of Convex-DS, a dataset to train intent detection procedures for explainable RS in the domain of hotels. This paper addresses our question RQ5.

Contributions as first author: I was responsible for the development of system functionalities required for collection and annotation of the dataset (including intent detection procedures, answer generation module, user interface and annotation application). I was also responsible for data collection and data annotation studies and procedures, data processing and analysis, and the writing of the paper. The design, methodologies and results of this study were discussed with Prof. Dr. Jürgen Ziegler, who also contributed extensively to the revision of the text.

2.6 Paper 6

Hernandez-Bocanegra, D.C. & Ziegler, J. (2021). Explaining Recommendations Through Conversations - Argumentative Dialog Model and Comparison of Interaction Styles. Manuscript under review in ACM Journal Transactions on Interactive Intelligent Systems.

In this paper, we extend work reported in Paper 3, Paper 4 and Paper 5. New content of this paper addresses the design and implementation of ConvEx, a conversational agent to provide natural language explanations, and the results of a user study to compare the interfaces: GUI navigation and natural language conversation, and their effect on users evaluation of RS. This paper addresses our questions RQ3, RQ4, RQ6 and RQ7.

Contributions as first author: I was responsible for the design of the dialog policy, and the quality testing of the conversational agent ConvEx. I developed the natural language understanding module of ConvEx, and the recommender system REST services published for ConvEx consumption. I was also responsible for the user study setup and execution, data analysis and the writing of the paper. The design, methodologies and results of this study were discussed with Prof. Dr. Jürgen Ziegler, who also contributed extensively to the revision of the text.

3 Conclusions and Future Work

3.1 Discussion and Design Guidelines

It is assumed that by developing intelligent systems able to explain their outcomes and decisions, users will have a better understanding of how those systems work, which in turn will lead to a greater confidence in the system by users (Miller, 2018; Pu et al., 2011; Tintarev & Masthoff, 2012; Walton, 2011). However, there is still a lack of widely adopted criteria of what makes an explanation a *good* explanation, which, as argued by Miller (2018), should not be left to the discretion of the intuition of intelligent system developers.

Throughout this dissertation, a series of designs for personalized explanations in review-based RS is introduced, which leverage the richness of content expressed in user reviews, to provide appropriate support for the various claims and premises presented by the RS. Particularly, we propose an approach to explanations as interactive argumentation, elaborated through an iterative, user-driven design process. Throughout this process, we detected different interface features and components that might contribute to a better perception of explanation quality. We also achieved a better understanding of explanation needs of RS users, in the context of goods with a strong experiential character, such as hotels, which we found differ from explanation needs of other types of users, e.g. artificial intelligence domain experts interacting with machine learning algorithms. The studies conducted throughout the dissertation advanced the understanding of user concerns related to RS transparency, and led to a proposal for interactive explanations that proved to have a positive impact on the evaluation of transparency and trust by users.

Our proposal of interactive explanations regards an explanation as a process, an exchange of questions and answers between system and user, as it would occur in a human-to-human conversation. This view is inspired by dialog models of conversation (e.g. Walton (2011)), and is aligned with the view of explanations in intelligent systems as a social process, as advocated by Miller (2018), who suggest taking into account perspectives to explanations from disciplines such as philosophy or psychology, when developing intelligent systems. Under this perspective, an explanation should be regarded as a "process of transferring knowledge between explainer and explainee"

(Miller, 2018), where the goal is that the explainee has sufficient information to understand the reasons that generate an event or a decision by a system. Miller (2018) draws four main characteristics of explanations from a social science perspective, which are usually overlooked by developers of intelligent systems, but which could make an important contribution to better evaluation of the quality of explanations by users: 1) "probabilities or statistics probably don't matter", 2) "explanations are selected", 3) "explanations are social", 4) "explanations are contrastive". In the following, our results are contrasted with these statements, and corresponding design guidelines are formulated, to be applied by review-based RS developers, for experience goods domains, e.g. hotel. Furthermore, the additional points are discussed: 5) need for procedural and algorithmic explanations and 6) relevance of detailed information about the inference of user's profile.

1. Statistics *do* matter, but are not enough. According to Miller (2018), providing statistical overviews might not be as efficient as providing support on real causes of an event. In this respect, our findings from Paper 1 showed that explanations showing an aggregated view of opinions using percentages of positive and negative opinions were perceived as more satisfying, compared to explanations that only provide a mere textual summary of opinions, suggesting that percentages may serve as easy anchors to convey more compelling information, while summaries may be perceived as too imprecise to convince. However, findings from Paper 3 and Paper 6 indicate that providing percentages is not enough, and that providing examples extracted from customer reviews that support such an overall view results in a better perception of explanation quality. We found this was more accentuated in users with a greater tendency to exhaustively search for information, a trait prevalent in more rational users during decision-making, and also in users who tend to take into account others' opinions. Additional findings reported in Paper 6 also indicate that users more familiar with information visualization perceived better explanations with the overall percentages rather than partial views of performance for a few aspects. Consequently, the following guideline is suggested:

Design guidelines

- Provide an initial overall view of statistics, either by listing aggregate values in a table, a bar chart, or embedded in short texts, where various aspects are listed, and allow users to go into detail on statistics that require further elaboration, taking advantage of interaction options.

2. Explanations should be selected. When providing an explanation, people tend to *select* a subset of possible causes of an event, instead of giving all the reasons that lead to such an event Miller (2018), due to cognitive reasons: full causal chains might be too extensive to process (Hilton, 2017). The above is aligned with our observations of Paper 2 that not all users regard every explanation component as useful, and of Paper 3 that not all users require to dig explanations into the same level of detail, which is highly dependent on differences in user characteristics (such as the way how information is processed when making decisions). Additionally, we argue that both context (e.g. the domain subject to recommendations) and the momentary needs by users also play a role in such explanatory needs.

It is important to note that, as reported in Paper 1, users did not report significantly better (static) explanations with a higher level of detail, while in Paper 3 and Paper 6 we did find a significant improvement in reported explanation quality when further level of detail was provided. Although seemingly contradictory, the difference that plays a crucial role is that explanations in the studies reported in Papers 3 and 6 were interactive, in contrast to the static explanations provided in study reported in Paper 1. Under our proposal of explanations as argumentative interaction, further details on why items are recommended are only offered in response to users' argumentation requests, as part of a question-answer exchange between system and user. Under this scenario, the system chooses to offer a set of additional causes only at the user's discretion, instead of presenting all possible causes in one single explanation. The above also aligns with guidelines by Hilton (1990), who argue that good explanations must be brief and answer the question asked, rather than presenting potentially irrelevant statements, which in turn aligns with Grice's maxims of conversation (Grice, 1975): quantity (statements should be informative but not more than required), relevance (get to the point and answer the question asked, not provide irrelevant answers), manner (be brief and avoid ambiguity). Thus we argue that developers can leverage interactive explanations, in order to find such a balance between explanations sufficient in content, but brief and relevant at the same time. Thus, and based in our findings, the following guideline is suggested:

Design guidelines

- Provide interactive options to users to access detailed explanatory information at will, instead of providing every possible cause of the system decision in a single step.
- Provide concise answers to users questions. In the case of aggregated customer opinions, visualizations using a table or a bar chart allow to provide a succinct and effective overall picture, that could benefit specially users more familiar with information visualization.
- When using a conversational agent to reply questions when no aspects are addressed (e.g. "why is option X good?"), provide answers involving aggregated opinions limited to a low number of aspects (the most relevant, according to the

preferences detected by the RS algorithm), and ask the user whether information on further aspects is still needed.

3. Explanations should be argumentative *and* dialogical. Miller (2018) argues that explanations are an inherently conversational process, where an "explainer transfer knowledge to an explainee", a view shared by proponents of dialog models of explanation (e.g. Walton (2011)), who argue that static argumentation may not be sufficient to achieve understanding by the explainee, as a dialectical approach to explanations may be. Miller refers to the "conversational" process, even warning that conversational does not necessarily imply a natural language conversation. This view is also shared by Jannach et al. (2020) and their definition of conversational RS (conversation in RS can be enabled by both traditional GUI functionalities and natural language interaction).

In turn, the view of explanations as a social process is aligned with our proposal of explanations as an interactive argumentation, and supported by our finding: providing explanations with a higher degree of interactivity through an argumentative conversation between the user and the system influences positively the evaluation of the RS by the users. Thus, the following guideline is suggested:

Design guidelines

- Provide explanations through a conversation between user and system. Such conversation can be enabled by using both GUI-based navigation controls (reflecting possible questions that the user might ask the system), or a conversational agent (where the user can state their questions or argument requests in their own words).
- Attempts at explanation by the system, as answer to users questions, should reflect an argumentative structure, consisting of statements supporting system claims (e.g. pros), as well as rebutting statements (e.g. cons). The RS should also include options for users to ask follow-up questions that contest such statements.

We note, however, that our proposed model for interactive explanations represents one among other possibilities for an explanatory dialog in RS. While our scheme depicts a "system explains first" alternative, future work should explore the adoption of a model in which the user asks the questions first, as well as validation of how much users' explanatory questions would vary under this scenario, compared to our dimension-based model of user intent. In addition, our model is designed in such a way that the questions are mainly asked by the user. However, future work could also explore the implications and effects of the system asking questions to the user, of the type: "would you like to get information on how recommendations were calculated?" or "do you want to know how your profile was inferred?". Here, we speculate that if the system takes a more proactive stance during the dialog, guiding the user on the type of queries that can be formulated, the user may decide accordingly to ask such questions, in contrast to the scenario where the user might be unaware of the system's capabilities, and therefore avoids asking certain types of questions (as further discussed in points 5 and 6 of this section).

4. Explanations should *not only* be contrastive. Authors such as Hilton (1990), Lipton (1990) and Miller (2018) argue that humans do not explain events' causes, but instead causes of why a certain event occurred and not another one, and that explanations basically reply to why-questions, which are actually of the form "why P rather than Q", even when Q is not explicit. Explanations of this form are then defined as *contrastive explanations*. Our findings suggest, however, that not all explanations in RS should be regarded as contrastive, and that the range of possible requests for explanations by users cover a much wider range than only comparative why-questions.

Contrastive explanations may be very common - almost prevalent - in typical machine learning classification problems (e.g. "Why is this picture classified as a dog, and not as a cat?"), while we argue that a distinction must be made for RS. In this case, predictions are not limited to a small subset of classes to compare as in prevalent explainable AI approaches, mostly related to machine learning classification outcomes. In RS, several "good options" are presented, and usually the predicted ranking is omitted. Thus, RS usually do not present a single, definitive recommendation that is better than all the others, but rather attempt to highlight aspects that may be attractive to users, for different options. So a direct comparison ("why X is better than Y?" or "between X and Y, which is best?") might be requested only after an initial evaluation of items at single-item level.

Such inference is aligned with our findings in Paper 4 and Paper 5: 1) questions regarding a single item outnumbered those with explicit tuples to compare, or no mention to items at all; 2) users do ask comparative questions, both by making explicit the items to be compared, and without such indication (i.e. superlative comparisons, in the form "which best"), but the number of non-comparative questions outnumbered the comparative questions; 3) factoid questions are the most prevalent type of questions, followed by subjective evaluation questions such as "how good is X", and why-recommended questions. This latter finding leads us to conclude that beyond contrastive, explanations should also be descriptive. That is, users not only expect the system to explain why an item is recommended in terms of how good it is with respect to others, but also in terms of the item's properties itself, as it would the case in content-based systems (e.g. "we recommend this camera because it has X resolution"), which does not necessarily imply a contrast with other items. In consequence, and based on our findings, the following guideline is suggested:

Design guidelines

- Provide explanations related to different types of assessment of an item: factoid, subjective evaluation, and why-recommended questions.
- To successfully answer the factoid questions, we suggest integrating alternative sources of information, beyond input used by the explanatory method, such as metadata or information from external sources (e.g. location, menus).
- Enable users to ask comparative questions, involving not only "why X instead of Y" kind of statements, but also superlative questions ("why X the best?").

Additionally, our findings lead to further considerations:

5. Procedural or algorithmic explanations have much less priority than item-based explanations. As reported in Paper 4 and Paper 5, we observed that users showed very little interest in aspects specifically related to the system or in details of how recommendations were calculated. In the context of our research, users showed strong interest in properties and reported opinions about items and their aspects, and how well items perform in comparison to other recommendations, rather than specifics on how the system got to such a decision. In this regard, we speculate that participants might already hold a mental model (i.e. the subjective notion that users have about how a system works (Norman, 1983)), that could have been transferred and applied to their interaction with our developed system. As discussed by Ngo et al. (2020), users tend to apply mental models when feasible, to deal with cognitive effort, i.e. a transfer of a mental model from one system to the other (Norman, 1983). Thus, instead of focusing on algorithmic details, which might be assumed to be transversal between similar RS, users might decide to focus on reasons backing the quality of the items, as inferred by the system.

We speculate on additional factors that may also have played a role in this observation. In the case of the interaction with a conversational agent, users may refrain from asking questions that they are not sure the system will answer (Jain et al., 2018). Also, it may be the case that lay users suspect that answers to algorithmic questions would reveal much more information than necessary, a factor that can lead to feelings of overwhelm for users (Ananny & Crawford, 2018). We also noted in studies reported in Papers 4 and 5 scarce further curiosity on the input used by the RS. This seems to be a consequence of the nature of the explanations we provided: they indicated that the recommendations were based primarily on reviews. The above became clear to users, as reported in Paper 6 (prevalent recognition of customer reviews as input for recommendations). Furthermore, results reported in Paper 2 indicate users found the indication on input used for recommendations as helpful, while this was not the case for details on inference of user profile, as discussed below. We note, however, that the implications described in this point are highly domain dependent, which is discussed further in the limitations section.

Thus, and based on our findings, the following guideline is suggested:

Design guidelines

- For domains related to experience goods, a strong focus on items properties and opinions reported for the item is preferred in explanations, rather than providing details on algorithmic procedures.

6. Detailed information about inference of one's own profile is not always relevant. As discussed in Paper 2, the quality of explanations was not reported as significantly better, when including detailed information about how user preferences were calculated (i.e. how often the user reported on aspects in their previous reviews), compared to explanations omitting such information. Moreover, as also reported in Paper 2, only a minority of users found this information useful. In this respect, we found that explanations without detailed information on user's profile inference were significantly easier to understand than those including such information. The above suggests that our proposed design to show details on users' profile inference and their alignment with customer-reported opinions (Fig. 1.5) is subject to improvement in future work. Such review should also consider the limitations of using the EFM algorithm, which calculates user preferences only based on the number of times the user commented on an aspect in their previous comments, which might be insufficient - even inaccurate - to form an adequate picture of user preferences, as discussed in detail in Paper 2, and later in the limitations of this dissertation.

It is important to note that, despite these limitations, the idea that providing detailed information on preference calculation may not be as useful to users in the context studied is reinforced by: 1) findings reported in Paper 4 and Paper 5: users expressed little curiosity about this aspect, with virtually no explicit user-generated questions about how their preferences were calculated, used or weighted to get recommendations. 2) findings reported in Paper 6 suggest that users could recognize to a fair extent that the recommendations were based on their preferences (in the case of GUI-based navigation), or the questions they formulated to the conversational agent (in the case of natural language interface).

The above seems, however, to be very specific to the type of domain we took as a case of study. Hotels are considered an experience good, where evaluation of alternatives and final decision is highly influenced by word-of-mouth. Here, the perspective and opinion of others might be more relevant than details about their own inferred profile. In addition, the hotel domain involves mostly a low risk of negative consequences associated with inaccurate preference calculation, which is not the case for domains where such risk may be considerably higher, such as medical or financial domains. In the latter case, a deeper understanding of how the self profile is calculated, and how it influences the final prediction of the system could indeed be relevant for users.

Notwithstanding the points discussed, our observations indicate that users do benefit from personalized explanations under our approach, i.e. those indicating that recommendations were generated on the basis of aspects relevant to them. Thus, the following guideline is suggested:

Design guidelines

- For domains related to experience goods, while it may not be necessary to reveal all the details of how user preferences are calculated, nor the weight they have in the final predictions, the RS must indicate in explanations that recommendations take into account aspects that are important to the user.

3.2 Limitations and outlook

Motivated interaction with the RS. The online nature of the user studies discussed in this dissertation involve a number of factors that are important to acknowledge. First, participants were recruited on crowdsourcing platforms where accounting for homogeneous demographics was not always possible. Also, promoting an effective execution of the task involves important challenges. Despite the measures we took after each iteration throughout the project (stricter attention controls, or bonuses based on answers to open-ended questions), these cannot beat a real motivation, e.g. choose a hotel that one actually needs. Therefore, future research needs to test the effects of the guidelines provided by this dissertation in real production web sites, for example, using A/B tests.

Generalization to other domains. Some of the findings of our studies could be generalized to other domains. For example, systems that collect customer reviews and recommend experience goods such as restaurants, could benefit users by providing explanations containing aggregate views of pros and cons for different aspects, that rely heavily on subjective opinions. Although shopping RS also collect and provide customer reviews as aids while making decisions, it is necessary to validate to what extent an explanation based mainly on subjective information competes with content-based explanations, which reflect more objective information on items' features. Consequently, future work could address the trade-off between review-based and content-based explanations, and possible paths of integration, given the potential benefits that both types of sources imply for the user experience.

In regard to the dataset we published for intent detection in dialog systems, ConvEx-DS, while it might also generalize properly to domains related to experience goods, further work is needed to test this inference. Here, recent advances in automatic natural language processing, and in particular, in transfer learning techniques, allow to obtain linguistic representations that can serve as a basis for training classifiers in similar domains. As of the intent model itself, the dimensions and its values do not involve domain concepts as such, so its generalization might be less problematic. Dimension *scope* refers to the number of entities referred in question (single, tuples or indefinite), which can be detected using name entity methods that are widely used in many domains; dimension *comparison* can be detected on the basis of comparison relations (e.g. "more than", "best of") rather than on domain specific features subject to comparison; dimension *detail* refers to the presence or absence of an aspect in a question, but it does not contain any indication on how these aspects should be detected or defined. As for dimension *evaluation*, future work needs to validate the distinction between subjective assessment and factual information, as different domains may differ in this regard.

User profile inference and disclosure. While in discussion (point 6) it is claimed that detailed information about inference of one's own profile might not be relevant for users, we acknowledge the following limitations in this respect. First, studies from Paper 2 were based on mockups, where no explicit elicitation of participants' preferences was done, although we mitigated this limitation in studies reported in Papers 3, 4, 5 and 6. Second, the use of the EFM method (Zhang et al., 2014) poses, in our view, an important constraint in regard to user profile inference: it only uses comments written by the user to it. Such information may not fully reflect the true preferences of the user. For example, one can talk about an aspect in a review, but not necessarily because it is very relevant to oneself, but because something was especially bad (or extraordinarily good) about it. Thus, future work needs to address the integration of review-based explanatory methods (such as EFM) with methods involving a more explicit elicitation of users' preferences, for example, critique RS methods.

Use of templates to explain. We used templates to generate explanations reporting on aggregated opinions (e.g. "because n% of customers reported positive comments on [aspect]"), in both GUI-navigation and natural language conversation interfaces; although for the latter we used a set of different statements for each intent type, so that slightly different answers could be generated every time the user express the same intent. Nevertheless, the use of template-based explanations may be perceived as too repetitive for users, while implementations based on natural language generation (NLG) may be better received as seemingly more flexible. Dialog approaches based on neural generative models can contribute in this regard, as their output may more closely resemble human responses as part of a dialog, as no templates are used. The above imply, however, important challenges: in one hand, dialog approaches based on neural models may require extensive datasets to train (Li et al., 2018). On the other hand, responses still need to report and reflect reasons consistent with the underlying explanatory method used. Therefore, as future work, we plan to analyze implications of extending our approach, to generate explanations that are not template-based, leveraging NLG techniques, but still reflecting an argumentative structure and the underlying RS method.

Flexibility of the dialog policy. As part of our developed conversational agent, we implemented a handcrafted dialog manager, where the state and policy were defined as a set of rules defined by the research team, and tied strictly to the explanatory process. Although adequate to test our approach to conversational explanations, future research should address more flexible approaches, to broaden the dialog policy to support related processes arising from interaction with a RS, e.g. preference elicitation, item search, customer service or booking. Considering a conversational approach that integrates these processes is relevant, since, in domains such as hotels, these needs would rarely occur in isolation. In this regard, end-to-end neural network approaches seem promising for this purpose, as they allow for a more dynamic modeling of possible dialogues with different goals, as proposed for example by Z. Liu et al. (2020).

Selection of customer reviews excerpts. Our approach involves providing excerpts of customer reviews as backing for aggregation-type explanations, filtered by aspect and sentiment. It is claimed in discussion (point 2) that explanations should be selected. Although relevance was taken into account in our implementation of interactive explanations (only comments on the specific aspect or feature requested by the users were presented), we did not implement methods for sorting customer comments according to their helpfulness or argumentative quality. Thus, future work needs to integrate techniques to determine the quality of customer comments, as well as to evaluate their effect, both on the accuracy of recommendation prediction and on the evaluation of the system by end-users.

Conversational agent can not reply to all questions. To date, creating dialog systems able to answer all possible users' questions remains unrealistic (Moore & Arar, 2018). Nevertheless, by using our approach, our implementation was able to generate answers to more than 80% of the users' questions, as reported in Paper 5, a number that we consider rather high. Addressing intent detection as a text classification problem, however, might be insufficient when dealing with questions that are too specific, particularly in regard to factoid questions. Therefore, to close the gap of questions that could not be answered by the system, future work may also consider leveraging information retrieval approaches, such as question and answer (QA), in addition to intent detection procedures, as per our proposal. In addition, future work should address the integration of information sources beyond customer reviews or hotels' metadata, in order to cover very specific explanatory needs of users, e.g. external location services. The above can contribute to answering questions involving item surroundings, distances to places of interest or transport means, which were very common concerns expressed in user studies reported in Papers 4, 5 and 6.

3.3 Conclusion

This dissertation presents a proposal for personalized explanations as interactive argumentation in review-based RS, inspired by dialog explanation models and formal argument schemes. This proposal allows users to go from aggregated customer opinions to detailed extracts of individual reviews, in order to facilitate a better understanding of the claims made by the RS. To this aim, a user-centered approach was followed, in order to guide and evaluate the design of the proposed explanatory interfaces, from a user perspective. This work reports empirical evidence, that providing a higher degree of interactivity in explanations contributed to a more positive evaluation of transparency and trust in the system, both in the GUI-navigation and natural language conversation variants of the RS implemented. Finally, this dissertation contributes to a better understanding on how to positively impact users' evaluation of explanation quality, system transparency and trust in the system, by providing guidelines on how to integrate interactive aspects to explainable RS interface design.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 chi conference on human factors in computing systems - chi 18* (p. 1–18).

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, *20*(3), 973–989.

Arioua, A., & Croitoru, M. (2015). Formalizing explanatory dialogues. *Scalable Uncertainty Management*, 282–297.

Bauman, K., Liu, B., & Tuzhilin, A. (2017). Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (p. 717–725).

Bentahar, J., Moulin, B., & Belanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, *33*(3), 211–259.

Berkovsky, S., Taib, R., & Conway, D. (2017). How to recommend?: User trust factors in movie recommender systems. In *Proceedings of the 22nd international conference on intelligent user interfaces* (p. 287–300).

Bilgic, M., & Mooney, R. J. (2005). Explaining recommendations: Satisfaction vs. promotion. In *Proceedings of the workshop on the next stage of recommender systems research, beyond personalization 05* (p. 13–18).

Blair, J. A. (2012). The possibility and actuality of visual arguments. *Tindale C. (eds) Groundwork in the Theory of Argumentation*, *21*, 205–223.

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, *36*(2), 3–10.

Buhalis, D., & Cheng, E. S. Y. (2020). Exploring the use of chatbots in hotels: Technology providers perspective. *Information and Communication Technologies in Tourism*, 231–242. doi: https://doi.org/10.1007/978-3-030-36737-4_19

Burack, J. A., Flanagan, T., Peled, T., Sutton, H. M., Zygmuntowicz, C., & Manly, J. T. (2006). Social perspective-taking skills in maltreated children and adolescents. *Developmental Psychology*, *42*(2), 207–217.

Burke, R. (2007). Hybrid web recommender systems. *The Adaptive Web, LNCS*, *4321*, 377–408.

Carenini, G., Cheung, J. C. K., & Pauls, A. (2013). Multi document summarization of evaluative text. In *Computational intelligence* (Vol. 29, p. 545-574).

Chandler, M. (1973). Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. *Developmental Psychology*, *9*(3), 326-332.

Chen, C., Zhang, M., Liu, Y., & Ma., S. (2018). Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 world wide web conference on world wide web. international world wide web conferences steering committee* (p. 1583–1592).

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, *19*(2), 25–35. doi: https://doi.org/10.1145/3166054.3166058

Chen, L., & Pu, P. (2014). Critiquing-based recommenders: survey and emerging trends. , *22*(1–2), 3085–3094.

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining kdd 16* (p. 815–824). doi: https://doi.org/10.1145/2939672.2939746

Collaborative for Academic Social and Emotional Learning. (2013). 2013 CASEL guide: Effective social and emotional learning programs - preschool and elementary school edition.

Costa, F., Ouyang, S., Dolog, P., & Lawlor, A. (2018). Automatic generation of natural language explanations. In *Proceedings of the 23rd international conference on intelligent user interfaces companion* (p. 57:1–57:2).

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653–665.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dong, R., Mahony, M. P. O., & Smyth, B. (2014). Further experiments in opinionated product recommendation. In *Case based reasoning research and development* (p. 110–124). Springer International Publishing.

Donkers, T., Kleemann, T., & Ziegler, J. (2020). Explaining recommendations by means of aspect-based transparent memories. In *Proceedings of the 25th international conference on intelligent user interfaces* (p. 166–176).

Donkers, T., & Ziegler, J. (2020). Leveraging arguments in user reviews for generating and explaining recommendations. *Datenbank-Spektrum*, *20*(2), 181–187.

Driver, M. J., Brousseau, K. E., & Hunsaker, P. L. (1990). The dynamic decision maker.

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). *Collaborative filtering recommender systems*. Now Publishers Inc.

Gedikli, F., Jannach, D., & Ge, M. (2014). How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, *72*(4), 367–382.

Gerani, S., Mehdad, Y., Carenini, G., Ng, R. T., & Nejat, B. (2014). Abstractive summarization of product reviews using discourse structure. In *Empirical methods in natural language processing* (Vol. 53, p. 1602–1613).

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Journal of Revenue and Pricing Management*, *23*, 1498–1512.

Grice, H. P. (1975). Logic and conversation. *Syntax and semantics 3: Speech acts*, *3*, 41-58.

Habernal, I., & Gurevych, I. (2017). Argumentation mining in user-generated web discourse. In *Computational linguistics 43* (Vol. 1, p. 125–179).

Hamilton, K., Shih, S.-I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of Personality Assessment*, *98*(5), 523–535.

Harms, J.-G., Kucherbaev, P., Bozzon, A., & Houben, G.-J. (2019). Approaches for dialog management in conversational agents. *IEEE Internet Computing*, *23*(2), 13-22. doi: https://doi.org/10.1109/MIC.2018.2881519

He, X., Chen, T., Kan, M.-Y., & Chen, X. (2015). Trirank: Review aware explainable recommendation by modeling aspects. In *Proceedings of the 24th acm international on conference on information and knowledge management* (p. 1661–1670). ACM.

Hellmann, M., Hernandez-Bocanegra, D. C., & Ziegler, J. (2022). Development of an instrument for measuring users' perception of transparency in recommender systems. In *Proceedings of the 6th humanize workshop.*

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 acm conference on computer supported cooperative work* (p. 241–250). ACM.

Hilton, D. J. (1990). Conversational processes and causal explanation. , *107*(1), 65–81.

Hilton, D. J. (2017). Social attribution and explanation. , 645-676.

Hu, Y.-H., Chen, Y.-L., & Chou, H.-L. (2017). Opinion mining from online hotel reviews: A text summarization approach. In *Information processing and management* (Vol. 53, p. 436–449).

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference dis* (p. 895–906). doi: https://doi.org/10.1145/3196709.3196735

Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2020). A survey on conversational recommender systems. In (p. 1–35). doi: abs/2004.00646

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language information applications. In *Transactions on office information systems* (Vol. 2, p. 26-41).

Kirby, J. R., Moore, P. J., & Schofield, N. J. (1988). Verbal and visual learning styles. *Contemporary Educational Psychology*, *12*(2), 169–184.

Klein, L. (1998). Evaluating the potential of interactivemedia through a new lens: Search versus experience goods. In *Journal of business research* (Vol. 41, p. 195–203).

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. In *User modeling and user-adapted interaction* (p. 441–504).

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2019). Personalized explanations for hybrid recommender systems. In *Proceedings of 24th international conference on intelligent user interfaces (iui 19)* (p. 379–390). ACM.

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 chi conference on human factors in computing systems* (p. 5686–5697).

Li, R., Kahou, S., Schulz, H., Michalski, V., Charlin, L., & Pal, C. (2018). Towards deep conversational recommendations. In *32nd conference on neural information processing systems, neurips 2018* (p. 9725–9735).

Lim, N. R., Saint-Dizier, P., & Roxas, R. (2009). Some challenges in the design of comparative and evaluative question answering systems. In *In proceedings of the 2009 workshop on knowledge and reasoning for answering questions - kraq 09* (p. 15–18). doi: https://doi.org/10.3115/1697288.1697292

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, *27*, 247-266.

Liu, Y., & Shrum, L. J. (2002). What is interactivity and is it always such a good thing? implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of Advertising*, *31*(4), 53-64.

Liu, Z., Wang, H., Niu, Z.-Y., Wu, H., Che1y, W., & Liu, T. (2020). Towards conversational recommendation over multi-type dialogs. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Loepp, B., Herrmanny, K., & Ziegler, J. (2015). Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *Proceedings of the 33rd annual acm conference on human factors in computing systems - chi 15* (p. 975–984).

Loepp, B., Hussein, T., & Ziegler, J. (2014). Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the 32nd annual acm conference on human factors in computing systems - chi 14* (p. 3085–3094).

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In S. Fussell & R. K. (eds.) (Eds.), *Recommender systems handbook* (p. 73-105).

Louvan, S., & Magnini, B. (2020). Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th international conference on computational linguistics* (p. 480–496).

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). A grounded interaction protocol for explainable artificial intelligence. In *Proc. of the 18th international conference on autonomous agents and multiagent systems, aamas 2019* (p. 1–9).

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. In *Information systems research* (Vol. 13).

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Mishra, A., & Jain, S. K. (2015). An approach for sentiment analysis of complex comparative opinion why type questions asked on product review sites. *Computational Linguistics and Intelligent Text Processing Springer LNCS*, *9042*, 257–271.

Moore, R. J., & Arar, R. (2018). Conversational ux design: An introduction. *Studies in Conversational UX Design*, 1–16. (Springer International Publishing) doi: https://doi.org/10.1007/978-3-319-95579-7_1

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

Muhammad, K. I., Lawlor, A., & Smyth, B. (2016). A live-user study of opinionated explanations for recommender systems. In *Intelligent user interfaces (iui 16)* (Vol. 2, p. 256–260).

Nelson, P. J. (1981). Consumer information and advertising. In *Economics of information* (p. 42–77).

Ngo, T., Kunkel, J., & Ziegler, J. (2020). Exploring mental models for transparent and controllable recommender systems: A qualitative study. In *Proceedings of the 28th acm conference on user modeling, adaptation and personalization umap 20* (p. 183-191).

Norman, D. A. (1983). *Some observations on mental models*. New York, NY, USA: In Mental Models, Dedre Gentner and Albert L. Stevens (Eds.). Psychology Press.

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model User-Adap*, *27*, 393-444.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. In *Foundations and trends in information retrieval* (Vol. 2, p. 1–135).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (p. 311–318). Association for Computational Linguistics.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth acm conference on recommender systems - recsys 11* (p. 157-164).

Quarteroni, S., & Manandhar, S. (2008). Designing an interactive open-domain question answering system. *Natural Language Engineering*, *15*(1), 73–95. doi: https://doi.org/10.1017/S1351324908004919

Rago, A., Cocarascu, O., Bechlivanidis, C., & Toni, F. (2020). Argumentation as a framework for interactive explanations for recommendations. In *Proceedings of the seventeenth international conference on principles of knowledge representation and reasoning* (p. 805–815).

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Schnotz, W. (2014). Integrated model of text and picture comprehension. In *The cambridge handbook of multimedia learning (2nd ed.)* (p. 72–103).

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174.

Sokol, K., & Flach, P. (2020). One explanation does not fit all: The promise of interactive explanations for machine learning transparency. , *34*(2), 235–250.

Tintarev, N. (2007). Explanations of recommendations. *Proceedings of the 2007 ACM conference on Recommender systems, RecSys 07*, 203–206.

Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, *22*, 399–439.

Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (p. 353–382). Springer US, Boston, MA.

Tkalcic, M., & Chen, L. (2015). Personality and recommender systems. *In Recommender Systems Handbook,*, 715–739.

Toulmin, S. E. (1958). The uses of argument.

Verberne, S., van der Heijden, M., Hinne, M., Sappelli, M., Koldijk, S., Hoenkamp, E., & Kraaij, W. (2013). Reliability and validity of query intent assessments: Reliability and validity of query intent assessments. *Journal of the American Society for Information Science and Technology*, *64*(11), 2224–2237.

Vig, J., Sen, S., & Riedl, J. (2009). Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on intelligent user interfaces* (p. 47–56). ACM.

Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. (2002). A survey of user-centered design practice. In *Proceedings of the conference on human factors in computing systems chi 2002* (Vol. 4, p. 471-478).

Walton, D. (2011). A dialogue system specification for explanation. , *182*(3), 349–374.

Walton, D., & Krabbe, E. C. W. (1995). *Commitment in dialogue: Basic concepts of interpersonal reasoning.* New York: State University of New York Press.

Wang, N., Wang, H., Jia, Y., & Yin, Y. (2018). Explainable recommendation via multitask learning in opinionated text data. In *Proceedings of the 41st international acm sigir conference on research and development in information retrieval, sigir 18* (p. 165–174).

Wegener, D. T., Petty, R. E., Blankenship, K. L., & Detweiler-Bedell, B. (2010). Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Consumer Psychology*, *20*, 5-16.

Wu, Y., & Ester, M. (2015). Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Eighth acm international conference on web search and data mining* (p. 153–162). ACM.

Xiao, B., & Benbasat, I. (2007). Ecommerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly*, *31*(1), 137-209.

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*, 104–120.

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, *14*(1), 1-101. doi: http://dx.doi.org/10.1561/1500000066

Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management* (p. 177–186). doi: https://doi.org/10.1145/3269206.3271776

Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., & Ma., S. (2014). Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international acm sigir conference on research and development in information retrieval* (p. 83–92).

Appendix

Paper 1

The following paper is reused from:

- Hernandez-Bocanegra, D.C., Donkers, T., & Ziegler, J. (2020). Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*, 219-225. ACM, New York, NY, USA. doi: https://doi.org/10.1145/3386392.3399302

Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations

Diana C. Hernandez-Bocanegra
University of Duisburg-Essen
Duisburg, Germany
diana.hernandez-bocanegra@uni-due.de

Tim Donkers
University of Duisburg-Essen
Duisburg, Germany
tim.donkers@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

Recommender systems have achieved considerable maturity and accuracy in recent years. However, the rationale behind recommendations mostly remains opaque. Providing textual explanations based on user reviews may increase users' perception of transparency and, by that, overall system satisfaction. However, little is known about how these explanations can be effectively and efficiently presented to the user. In the following paper, we present an empirical study conducted in the domain of hotels to investigate the effect of different textual explanation types on, among others, perceived system transparency and trustworthiness, as well as the overall assessment of explanation quality. The explanations presented to participants follow an argument-based design, which we propose to provide a rationale to support a recommendation in a structured way. Our results show that people prefer explanations that include an aggregation using percentages of other users' opinions, over explanations that only include a brief summary of opinions. The results additionally indicate that user characteristics such as social awareness may influence the perception of explanation quality.

CCS CONCEPTS

• Information systems → Recommender systems; **• Human-centered computing → User studies**.

KEYWORDS

Recommender systems, user study, explanations

ACM Reference Format:
Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3386392.3399302

1 INTRODUCTION

Providing explanations of the rationale behind a recommendation can bring several benefits to recommender systems (RS). In particular, explanations may serve the following aims [27]: transparency (the system explains how it works), effectiveness (user can make good decisions), efficiency (user can make decisions faster), and trust in the system. Explanations based on collaborative filtering inform that a recommendation is based on preferences of similar users or items that the user liked in the past, e.g. Amazon's "Customers who bought ... also bought...", while content-based explanations present users with item features that can be relevant to them, e.g. [15, 29]. On the other hand, exploiting user reviews has drawn research interest recently, in particular to facilitate the generation of textual explanations, as proposed by [34] and [18], where a brief assessment of relevant aspects based on opinions from reviews is provided as explanation. However, avid users in need of specific details may be more satisfied when more robust arguments or a higher level of justification is provided. Here, important questions are still unresolved: Do users prefer concise explanations over those that include more specific details? Do they prefer an aggregated view of other users' opinions, over reading individual reviews written by similar users? Specifically, we regarded three different types of review-based explanation:

- Explanations with aggregated results: An accumulated view using bullet points and percentages of positive and negative opinions, as proposed by [11].
- Explanations with only textual summary: Summarization of opinions without bullet points nor percentages. It resembles a system generated review, as proposed by [8], and [3].
- Explanations using a helpful review: Indicate that the recommendation was based on the reviews that might be helpful to the user, as proposed by [6], and show just one of them as an example.

In this respect, both summaries and helpful reviews have proven to be an effective means of assisting users in making purchasing decisions, while helping them cope with the overwhelming amount of information available [3, 12, 16, 20, 23]. However, little is known about the suitability of one information style over another when offered as part of an explanation. Additionally, we also aimed to evaluate the effect of two levels of justification:

- High: Specific details about the main aspects (e.g. cleanliness) and finer-grained aspects (e.g. cleanliness of bathroom) are provided.
- Low: Only brief information about main aspects is provided.

UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-7950-2/20/07...$15.00
https://doi.org/10.1145/3386392.3399302

In addition, and taking into account that differences in user characteristics also contribute to differences in the general perception of RS [17, 31], we set out to focus on one of the main objectives of RS, which is to help users make better decisions. Particularly, individual differences between decision-making styles are determined significantly by preferences and abilities to process available information [9]. Accordingly, decision making styles are defined by [14] as a "habit-based propensity" to exhaustively search for information and to systematically evaluate possible alternatives (rational style), or to use a quick process based on hunches and feelings (intuitive style), in order to make decisions. Two main aspects provide basis to describe the differences in decision styles: information use (amount of information used during the process) and focus (alternatives addressed) [9]. In this respect, "good enough" information might be sufficient for some people, whereas others prefer to obtain and address all relevant information in order to minimize risks.

Additionally, we were interested in a second factor that may influence the way users perceive explanations: the extent to which they are able to adopt the perspective of others when making decisions. The rationale for this interest stems from the tendency of individuals to adjust their own opinions using those of others, while choosing between various alternatives [26], which may even be beneficial [32]. Particularly, individuals with greater perspective-taking skills tend to understand the views of others better [2, 5], skills that are also characterized by [10] as “social awareness”, which represents the propensity of individuals to empathize and take into account the opinions of others.

Accordingly, we aimed to answer the following research questions, in relation to our variables of interest (i.e. quality of explanation, transparency, effectiveness, efficiency and trust), and taking the hotels domain as a case in point:

RQ1: Does the *type* of explanation influence the perception of the variables of interest?

RQ2: Does the *level* of justification influence the perception of the variables of interest?

RQ3: Do individual differences in decision making styles and social awareness influence the perception of the variables of interest, when different types of explanation or levels of justification are provided?

Consequently, we conducted a study, in which users were asked to examine and read the explanations of a fixed set of hotel recommendations, and to report their perception of the quality of such explanations, as well as their perceived transparency, effectiveness, efficiency and trust of the system. The hand-made explanations provided were based on designed templates that follow principles of argumentation theory, as elaborated in detail in section 3.

2 RELATED WORK

Textual explanations in RS seek to provide reasons behind a recommendation, while assisting users making a decision. In this respect, in recent years there has been a growing interest in exploiting user reviews, given their richness in explanatory and argumentative information. [34] proposed a matrix factorization model to align explicit features and the latent representations of items and user preferences obtained from reviews, which allows to generate textual explanations based on templates (e.g. “You might be interested in [feature], on which this product performs well”). An extension of this work was presented by [6], who argued that reviews should have different weights when calculating predictions, and that, therefore, the most useful for the user should have a higher priority, and be used to generate explanations; however, no explanations are actually generated, but only selected reviews are provided. On the other hand, [8] proposed an natural language generation (NLG) procedure for creating reviews (as a real user would) and providing them as explanations without using templates, whereas [7] proposed a denoising mechanism to extract relevant sentences with explainable purposes, to generate natural language textual explanations (e.g. “The bottle is very light and the smell is very strong”). Additionally, [21] had proposed a series of interface variations, that provide users with display pros and cons scores using bars, as well as a report of feature performance in comparison with other alternatives; however, their visualizations do not provide details on the fine-grained aspects, nor possible reasons for conflicting opinions.

The above approaches result in explanations that may be perceived by users as being too general, and lacking solid arguments to justify the recommendation offered. On the other hand, [4] proposed a framework to generate arguments in the context of tasks like selecting a house to buy. [33] compares explanations with brief sentences and an argumentative structure - two facts and a claim -, for recommendations of hiking routes, energy and mobile phone plans; however, no counter-arguments are provided. [18] proposed a method based on [1] for generating explanations with convincing arguments in a mobile shopping recommender using templates: strong argument (e.g. “Mainly because you currently like X.”), supporting argument (e.g. “Also, slightly because of your current interest in X.", and negative argument (e.g. “However, it has the following features you don’t like: X, Y (...).”). The rather concrete and brief sentences proposed by [1] and [18] are oriented to provide interactive explanations in the mobile domain, where users might face both space and time limitations. However, we aimed to investigate the effect that more detailed explanations may have on users’ perception of recommender systems, while keeping an argumentative nature. To this end, we propose an explanations design with an argumentative structure, that is inspired on the scheme proposed by [13], a variation of original Toulmin’s model [28], that seeks to represent the kind of arguments usually provided in user-generated web discourse.

3 EXPLANATION DESIGN

We designed a series of templates that represent the combination of the two factors: *type* of textual explanation and *level* of justification. These templates were used to create the explanations we presented to participants in our empirical study. Table 1 shows the designed templates. Furthermore, the proposed design reflects an argumentative structure, inspired by the scheme proposed by Habernal et al. [13], and includes: a conclusion that informs how good the choice is for the user, evidence that supports such a claim, and possible reasons behind contradictory opinions.

Additionally, we considered a number of template variations in order to explain items with higher prediction ratings (*very good* or an *adequate* option), or lower prediction ratings (*not so good* option), depending on whether positive opinions are much much greater (very good) or greater (adequate) than negative ones, or if they are more negative than positive (not so good). These variations are represented mainly by differences in the rebuttal and the backing section of the explanation, as well as the presence of refutation statements, as depicted in the scheme of figure 1.

Explanations with aggregated results: Summarizes opinions found in reviews using bullet points and percentages of positive and negative opinions. It corresponds to the "Aggregation" condition of the empirical study.

Explanations with only textual summary: Summarizes opinions using just text (no bullet points nor percentages). The condition "Summary" refers to this type of explanation.

Explanations using a helpful review: This type of explanations indicate that recommendation was based on information provided in helpful reviews, and offers one of such reviews as an example. The condition "Review" refers to this type of explanation.

In turn, every type of explanation is provided in one of two variations:

Low level of justification. To address the main aspects of interest to users (e.g. overall cleanliness), without further elaboration or details.

High level of justification. To address the main aspects of interest to users (e.g. overall cleanliness) by providing fine grained details with several sentences about more specific aspects (e.g. cleanliness of bathroom).

4 EMPIRICAL STUDY

We intended to compare the users' subjective assessment of different types of explanation and different levels of justification. In particular, we hypothesized the following:

H1: People will be more satisfied with explanations that involve a higher level of justification.

H2: People will be more satisfied with aggregated explanations as opposed to mere summaries.

H3: People will be more satisfied with explanations that involve helpful reviews as opposed to mere summaries.

H4: More rational users would prefer a higher level of justification and explanations that involve helpful reviews or an aggregation of opinions, as opposed to summaries.

To test the above, we recruited 152 participants (87 female, mean age 39.84 and range between 18 and 75) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 95%. Although 334 workers completed the task, only 152 workers passed the quality check (i.e. at least 6 of the 7 validation questions were answered correctly, more than 20s were spent on the recommendation step and more than 30s on the evaluation questionnaire), so only the data for these participants were used for the following analysis. This sample size allows us to achieve a statistical power of 82.5% with the performed MANCOVA analysis (α =0.05). Participants were rewarded with $0.8 (time to complete task in minutes: M=8.56, SD= 1.86)

The study follows a 3x2 between-subjects design, and each participant was assigned randomly to one of six conditions that represent the combination of the two factors: *type* of explanation and *level* of justification. Participants were presented with a prototype that provided them with a fixed list of 5 hotels that represented the recommendations for a hypothetical hotel search. Each recommendation included an explanation of why the item was recommended. After the participants explored the information for all the hotels, they were asked to rate their perception of the recommender and its explanations. No real system was used to generate recommendations or explanations, as the main objective here was to test users' perception of explanation design.

Conditions: We regarded three different types of explanation: with aggregated results ("aggregation"), with only textual summary ("summary") and explanations using a helpful review ("review"). We also evaluated the effect of two levels of justification: "high" and "low". Section 3 provides further details on every type and level.

Procedure: After some questions on demographics, users answered the questionnaire on user characteristics. Instructions to participants indicated that a list of 5 hotels would be displayed, representing the results of a hypothetical search for hotels already performed. Here, participants were instructed to click the button "View Details" of each hotel and read the information provided, including the explanation of why the item was recommended. We then presented a cover story, which sought to establish a common starting point in terms of travel motivation (a business trip), and the presumed aspects of greatest interest to the user (cleanliness and location). The cover story also stated that different recommended hotels within the same price range would be shown. The users were then presented with a list of recommended hotels and their explanations. An example of the functionality provided to the users is shown in figure 2. The list of hotels, hotel names, photos, prices and ratings were the same for all users. Only the explanations provided varied according to the condition to which each participant was assigned. Next, users answered the evaluation questionnaire. In addition, we included an open-ended question, so that participants could indicate in their own words their general opinion about the explanations provided. We included 4 validation questions to check attentiveness within the questionnaires, and 3 validation questions related to the content of both textual and visual elements presented throughout the task.

Questionnaires:

User characteristics: We used the Rational and Intuitive Decision Styles Scale [14], and the scale of the social awareness competency, proposed by [10]. We used a 1-5 Likert-scale to evaluate all the items (1:Strongly disagree, 5: Strongly agree).

Evaluation: We used items from: [25] to measure the perception of transparency, [17] of effectiveness, [19] of efficiency, and [19] of trust. Finally, we also adapted 3 items from [17] to address explanation quality. We used a 1-5 Likert-scale to evaluate all the items (1:Strongly disagree, 5: Strongly agree).

5 RESULTS

User characteristics scores. In regard to decision making styles, we calculated the rational (M = 4.31, SD= 0.52) and the intuitive (M = 2.72, SD= 0.83) scores for each individual as the average of the values reported for the five items on both rational and intuitive decision-making style subscales. Likewise, we calculated the social awareness score (M = 3.99, SD= 0.49) for each individual based on

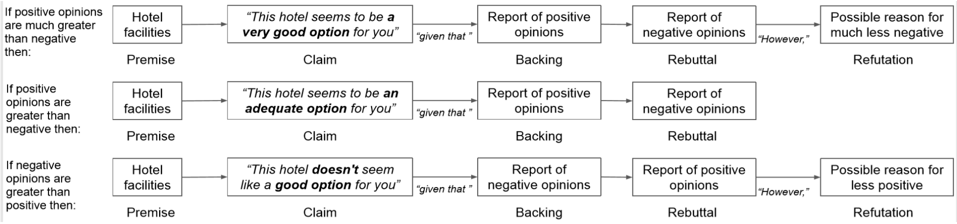
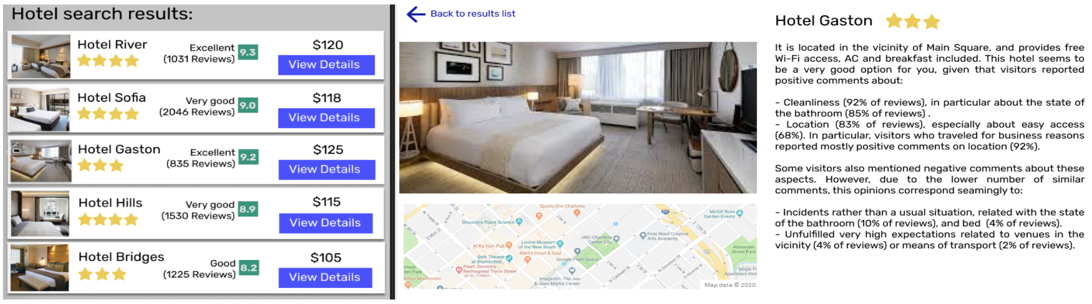

If positive opinions are much greater than negative then:
Hotel facilities
"This hotel seems to be a very good option for you"
"given that"
Report of positive opinions
Report of negative opinions
"However,"
Possible reason for much less negative
Premise
Claim
Backing
Rebuttal
Refutation
If positive opinions are greater than negative then:
Hotel facilities
"This hotel seems to be an adequate option for you"
"given that"
Report of positive opinions
Report of negative opinions
Premise
Claim
Backing
Rebuttal
If negative opinions are greater than positive then:
Hotel facilities
"This hotel doesn't seem like a good option for you"
"given that"
Report of negative opinions
Report of positive opinions
"However,"
Possible reason for less positive
Premise
Claim
Backing
Rebuttal
Refutation

Figure 1: Argument scheme used to create explanation templates, to provide reasons for recommending items with higher prediction ratings (very good or an adequate option), or lower prediction ratings (not so good option).

Aggregation	Summary	Review
Explanation Beginning (Both levels): [It is located in ..., and provides ... in all rooms]$_{\text{Premise}}$. [This hotel seems to be a very good option for you]$_{\text{Claim}}$, given that:	[It is located in ..., and provides ... in all rooms]$_{\text{Premise}}$. [This hotel seems to be an adequate option for you]$_{\text{Claim}}$, given that	[Based on the reviews that contain useful information and might be relevant to you.]$_{\text{Backing}}$, we believe that [this hotel is an adequate option for you]$_{\text{Claim}}$. This is an example of one of these reviews:
Low level: [n% of visitors reported positive comments about ... and n% about ...]$_{\text{Backing}}$. [Some visitors mentioned negative comments about ... (n%)]$_{\text{Rebuttal}}$, however [such claims are seemingly related to particular incidents, rather than a usual situation, or perhaps to very high expectations that were not met.]$_{\text{Refutation}}$	[usually ... is not a problem here, and the ... is ...]$_{\text{Backing}}$. [Although some reviews include negative comments about ...]$_{\text{Rebuttal}}$, [such claims seem to be more related to incidents rather than a usual situation, or perhaps to very high expectations that were not met.]$_{\text{Refutation}}$	["I've visited the River Hotel for a business trip. Coffee and tea in the room, clean, good location, near to ... Overall, a very good option, I would definitely come back!!!"]$_{\text{Backing}}$
High level: ... [visitors reported positive comments about: - The ... (n% of reviews), in particular about the state of ... (n% of reviews) - The ... (n% of reviews), especially about ... (n% of reviews).]$_{\text{Backing}}$. [Some visitors also mentioned negative comments about these aspects]$_{\text{Rebuttal}}$. However, [due to the lower number of similar comments, this opinions correspond seamingly to: - Incidents rather than a usual situation, related to the state of ... (n% of reviews). - Unfulfilled very high expectations related to ... (n% of reviews) or ... (n% of reviews).]$_{\text{Refutation}}$	[usually ... is not a problem here, in particular the state of ..., and the ... is quite good in general. The ... very convenient for your purposes, since it is ..., and it is also ...]$_{\text{Backing}}$. [Although some reviews include negative comments about ..., in particular in relation to ...]$_{\text{Rebuttal}}$, [such claims seem more related to incidents rather than a usual situation, or perhaps to very high expectations that were not met.]$_{\text{Refutation}}$	["I stayed at the Sofia Hotel in June. The location is convenient to... And very convenient when you need to work and not being disturbed by kids or drunk teenagers! My room was clean but more care for windows wouldn't hurt. Also, I think left the towels ..., I expected them to be changed, but that didn't happen until ..., but overall a minor issue, given the overall quality of the room. Parking is free, but you may not need it, as ... Overall, you get what it is advertised. I'd come back"]$_{\text{Backing}}$



Hotel search results:

Hotel River ★★★★	Excellent (1031 Reviews)	9.3	$120 View Details
Hotel Sofia ★★★★	Very good (2046 Reviews)	9.0	$118 View Details
Hotel Gaston ★★★	Excellent (835 Reviews)	9.2	$125 View Details
Hotel Hills ★★★★	Very good (1530 Reviews)	8.9	$115 View Details
Hotel Bridges ★★★	Good (1225 Reviews)	8.2	$105 View Details

Hotel Gaston ★★★

It is located in the vicinity of Main Square, and provides free Wi-Fi access, AC and breakfast included. This hotel seems to be a very good option for you, given that visitors reported positive comments about:

- Cleanliness (92% of reviews), in particular about the state of the bathroom (85% of reviews) .
- Location (83% of reviews), especially about easy access (68%). In particular, visitors who traveled for business reasons reported mostly positive comments on location (92%).

Some visitors also mentioned negative comments about these aspects. However, due to the lower number of similar comments, this opinions correspond seamingly to:

- Incidents rather than a usual situation, related with the state of the bathroom (10% of reviews), and bed (4% of reviews).
- Unfulfilled very high expectations related to venues in the vicinity (4% of reviews) or means of transport (2% of reviews).

Figure 2: Prototype screens displayed in empirical study. List of recommended hotels (left) and hotel details of the 3rd hotel of the list (right), depicting an explanation of the aggregation type with a high level of justification. Location screenshot Map data ©2020

the values reported for the items of this scale. Figure 3a depicts the different scores distributions.

Evaluation scores. We calculated evaluation scores for every variable of interest (explanation quality and the explanations aims: transparency, effectiveness, efficiency, and trust), as the average of the individual values reported for the items corresponding to each variable. Table 2 show the descriptive evaluation results by type and level, respectively.

Analysis of covariance Since our dependent variables are correlated (see Table 2), we performed a MANCOVA analysis to evaluate the simultaneous effect of type of explanations and level of justification on all variables that represent user's perception, and to what extent the individual decision-making styles or social awareness might influence such perception. Here, evaluation scores were used as the dependent variables, *level* and *type* as fixed factors (independent variables), and user characteristics scores as covariates. Smaller ANCOVA analyses were also performed, to test the interactions between independent variables and covariates, and their effect on each of the dependent variables. The results are summarized below.

Multivariate effects:

Significant multivariate effects were found for the variables: type *F(5, 140) = 4.68, p <.001* and social awareness *F(5, 139) = 2.41, p <.05.* No significant overall effects were found for the level of justification, nor for the rational or intuitive decision-making style.

Univariate effects:

We performed a set of 5 ANCOVA analyses, to test interaction and main effects of the variables that reported a significant overall effect (type and social awareness) on each of the 5 dependent variables (explanation aims). Tests were conducted using Bonferroni adjusted alpha levels of .01 per test (.05/5)

Explanation quality: The type of explanation influences significantly the perception of explanation quality, *F(2, 146) = 5.37, p<.01.* A post-hoc test using Tukey HSD reveals a significant difference between aggregation and summary conditions (*p <.01*), such that the average explanation quality was significantly higher for aggregation (*M* = 3.98, *SD* = 0.65) than for summary (*M* = 3.56, *SD* = 0.75). No significant interaction was found between social awareness and type after the Bonferroni correction, *F(2, 146) = 5.37, p=.019*; however, we observed that the relationship between social awareness and explanation quality has a positive tendency for the aggregation and summary types, (aggregation having a steeper slope), whereas for review the relationship tends to be negative (Figure 3b).

Transparency: We observed that the type of explanation influences significantly the perception of transparency ("the system explains why the items were recommended"), *F(2, 146) = 5.49, p <.01.* A post-hoc test using Tukey HSD reveals a significant difference between aggregation and review conditions (*p <.05*), such that the average perception of transparency was significantly higher for aggregation (*M* =4.05, *SD* =0.69) than for review (*M* = 3.68, *SD* = 0.63). However, no significant influence of type was found in relation to whether users actually understood why the system recommended the items. There was also no significant interaction between type and social awareness, although a significant effect of social awareness on transparency was found, *F(1, 146) = 7.15, p <.01.* Here we observed a positive trend in the relationship between social awareness and transparency, as depicted in figure 3c.

Effectiveness: No main effects of type were found, neither significant interaction between social awareness and type.

Efficiency: No main effects of type were found, neither significant interaction between social awareness and type.

Trust: A significant effect of social awareness on trust was found, *F(1, 146) = 11.92, p<0.001.* Here we observed a positive trend in the relationship between social awareness and trust, as depicted in figure 3c. We found no major effects of type, nor significant interaction between type and social awareness.

6 DISCUSSION

We observed that the type of explanation seems to significantly influence the quality perception of explanations. Explanations that include an aggregated view with percentages of positive and negative opinions are perceived as more satisfying over explanations that only provide a mere summary of opinions, which confirms our hypothesis H2. This suggests that percentages may serve as easy anchors to convey more compelling information, while summaries may be perceived as too imprecise to convince. In fact, judgments and decision making can be influenced by changes in attitude, which in turn can result from the effortless use of cues such as numerical anchors, when people lack motivation or ability [24, 30]. In addition, although the difference in perception of quality between explanations with summaries and helpful reviews is not significant to confirm our H3 hypothesis, there seems to be a tendency to prefer reviews over summaries. This may reflect that some people trust a single opinion more than summaries that may hide details of special interest to them. Furthermore, there is not enough evidence to confirm our H1 hypothesis that users would prefer a higher level of justification in explanations, nor that reporting additional details of fine-grained aspects may influence the general perception of the recommender system. On the other hand, and contrary to our H4 hypothesis, we found no influence of rationality on this perception. First, it is difficult to make assumptions with respect to this variable since our sample is very skewed: to the right for the rational style and to the left for the intuitive, as depicted in Figure 3a. Additionally, this may be related to our observation that rationality and intuition are not diametrically opposed constructs: although most participants consider themselves to be someone who thoroughly evaluate available information, many of them also have a tendency to use their intuition when making decisions. In this regard, [14] have indicated that people with a greater tendency to process information in a rational manner (i.e. a prevalent rational cognitive style, according to [22]) are less likely to be intuitive decision-makers, whereas subjects with a greater tendency to process information in a more intuitive manner (i.e. a predominantly intuitive cognitive style) may be either rational or irrational decision-makers [14]. On the other hand, and although the interaction between social awareness and type of explanation is not statistically significant, we observed a tendency to prefer aggregated explanations for subjects with higher social awareness scores. A similar effect is observed for reviews, although smaller; here, summaries may sound too detached from actual opinions people express, therefore, the effect is negative for more social aware people.

In terms of transparency, the results suggest that, even when some types of explanations seem to serve better than others to

Table 1: Mean values and standard deviations of perception on explanation aims, per *level of justification* and *type of explanation* (n=152); values reported with a 5-Likert scale; high values of means represent a positive perception of recommender and explanations. Pearson correlation matrix, p<0.001 for all correlation coefficients.

Level:	Low		High		*Type:* Aggregation		Summary		Review		*Corr:* Variable			
Variable	M	SD	M	SD	M	SD	M	SD	M	SD	1	2	3	4
1. Explanation Quality	3.79	0.70	3.83	0.73	3.98	0.65	3.56	0.75	3.88	0.68				
2. Transparency	3.93	0.69	3.92	0.69	4.05	0.69	3.99	0.69	3.68	0.63	0.41			
3. Effectiveness	3.87	0.72	3.85	0.70	3.95	0.70	3.69	0.73	3.93	0.68	0.82	0.48		
4. Efficiency	3.96	0.78	3.90	0.79	4.07	0.65	3.70	0.93	4.00	0.71	0.58	0.45	0.70	
5. Trust	3.84	0.58	3.71	0.68	3.85	0.66	3.65	0.68	3.80	0.57	0.75	0.50	0.80	0.73

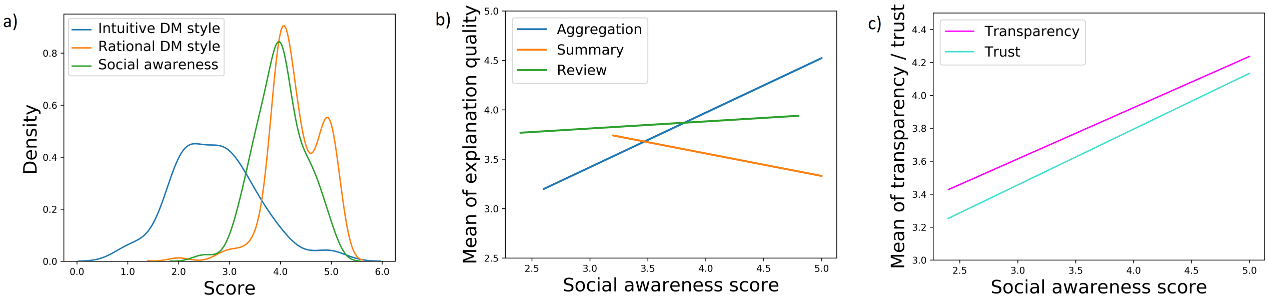


Figure 3: a) Kernel density estimate of user characteristics scores: rational and intuitive decision making styles and social awareness. b) Interaction plot for explanation quality (fitted means of individual scores) between type and social awareness. c) Effect of social awareness on transparency and trust (fitted means of individual scores). All scores on a 5-Likert scale.

explain the recommended items (in particular aggregations are perceived as more transparent than explanations based on helpful reviews), the users' understanding of the reasons behind the recommendations is not statistically different between types, i.e. a possible dichotomy between "the system explains why" and "I understood why". In this regard, some users mentioned, for example, that despite explanations were good, more details about how the algorithm actually works could further improve their understanding of reasons behind recommendations.

Finally, our results suggest that social awareness may play a role in the perception of both transparency and trust by users, that is, people with a higher disposition to listen and take into account others' opinions, tend to perceive the system as more transparent and to trust more in the recommender when the proposed explanations are provided, independent of their type or justification level.

7 CONCLUSION

In this paper, we have proposed the design of argumentative textual explanations, as well as examined and discussed the differences between types of explanations and levels of justification, their influence on users' perception of different characteristics of the system, and the influence that individual differences (namely decision-making style and social awareness) may have on such perception. We conclude that providing arguments based on aggregated results seems to be a meaningful way of presenting explanations. We cannot state though whether high or low levels of justification are *per se* better, or that differences between users' decision-making style influence signifficantly the perception of the proposed explanations. However, when taking into account another user characteristic, i.e. social awareness, differences in perception between users can be better understood, which can lead to better explanation designs and interaction possibilities. We believe that our findings lead to practical implications, e.g. that effective explanations should provide an initial aggregated overview of the main findings, and then allow the user to examine them in as much detail as preferred (e.g. by reading a list of the most useful reviews).

It is important, however, to recognize the limitations that the implementation of the proposed approach may have. For example, template-based explanations may be perceived as too repetitive for users, while implementations based on natural language generation (NLG) may be better received as seemingly more flexible. Therefore, as future work, we plan to extend our approach to the generation of explanations that are not template-based, leveraging NLG techniques, but still reflecting an argumentative structure. In addition, our evaluation has limitations, such as the use of a prototype instead of a system with real recommendations, as well as the use of Amazon Mechanical Turk, where despite our quality control implemented, it is difficult to encourage users to genuinely make a decision, which could guarantee higher quality in the execution of the task. Therefore, an evaluation on a real set and using a more effective motivation strategy will be part of the future work.

ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

REFERENCES

[1] Roland Bader, Wolfgang Woerndl, Andreas Karitnig, and Gerhard Leitner. 2012. Designing an explanation interface for proactive recommendations in automotive scenarios. In *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization (UMAP'11)*. 92–104.

[2] Jacob A. Burack, Tara Flanagan, Terry Peled, Hazel M. Sutton, Catherine Zygmuntowicz, and Jody T. Manly. 2006. Social Perspective-Taking Skills in Maltreated Children and Adolescents. *Developmental Psychology* 42, 2 (2006), 207–217.

[3] Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2013. Multi document summarization of evaluative text. In *Computational Intelligence*, Vol. 29. 545–574.

[4] Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. In *Artif. Intell.*, Vol. 170. 925–952.

[5] Michael Chandler. 1973. Egocentrism and Antisocial Behavior: The Assessment and Training of Social Perspective-Taking Skills. *Developmental Psychology* 9, 3 (1973), 326–332.

[6] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee.* 1583–1592.

[7] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2019. Generate Natural Language Explanations for Recommendation. In *Proceedings of SIGIR 2019 Workshop on ExplainAble Recommendation and Search (EARS19)*.

[8] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic Generation of Natural Language Explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion.* 57:1–57:2.

[9] Michael J. Driver, Kenneth E. Brousseau, and Phil L. Hunsaker. 1990. The dynamic decision maker. (1990).

[10] Collaborative for Academic Social and Emotional Learning. 2013. 2013 CASEL guide: Effective social and emotional learning programs - Preschool and elementary school edition. (2013).

[11] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive Summarization of Product Reviews Using Discourse Structure. In *Empirical Methods in Natural Language Processing*, Vol. 53. 1602–1613.

[12] Anindya Ghose and Panagiotis G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Journal of Revenue and Pricing Management* 23 (2011), 1498–1512.

[13] Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. In *Computational Linguistics 43*, Vol. 1. 125–179.

[14] Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The Development and Validation of the Rational and Intuitive Decision Styles Scale. *Journal of Personality Assessment* 98, 5 (2016), 523–535.

[15] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.* ACM, 1661–1670.

[16] Ya-Han Hu, Yen-Liang Chen, and Hui-Ling Chou. 2017. Opinion mining from online hotel reviews: A text summarization approach. In *Information Processing and Management*, Vol. 53. 436–449.

[17] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. In *User Modeling and User-Adapted Interaction.* 441–504.

[18] Béatrice Lamche, Ugur Adigüzel, and Wolfgang Wörndl. 2012. Interactive explanations in mobile shopping recommender systems. In *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE'14), held in conjunction with the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP'14).* 92–104.

[19] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. In *Information Systems Research*, Vol. 13.

[20] Susan M. Mudambi and David Schuff. 2010. What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly* (2010), 185–200.

[21] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Intelligent User Interfaces (IUI 16)*, Vol. 2. 256–260.

[22] Rosemary Pacini and Seymour Epstein. 1999. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology* 76 (1999), 972–987.

[23] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval*, Vol. 2. 1–135.

[24] Richard E. Petty and John T. Cacioppo. 1986. *Communication and persuasion: Central and peripheral routes to attitude change.* Springer-Verlag, New York.

[25] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems - RecSys 11.* 157–164.

[26] Janet A. Sniezek and Timothy Buckley. 1995. Cueing and Cognitive Conflict in Judge Advisor Decision Making. *Organizational Behavior and Human Decision Processes* 62, 2 (1995), 159–174.

[27] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22 (2012), 399–439.

[28] Stephen E. Toulmin. 1958. The Uses of Argument. (1958).

[29] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent User Interfaces.* ACM, 47–56.

[30] Duane T. Wegener, Richard E. Petty, Kevin L. Blankenship, and Brian Detweiler-Bedell. 2010. Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Consumer Psychology* 20 (2010), 5–16.

[31] Bo Xiao and Izak Benbasat. 2007. ECommerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly* 31, 1 (2007), 137–209.

[32] Ilan Yaniv and Maxim Milyavsky. 2007. Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* 103 (2007), 104–120.

[33] Markus Zanker and Martin Schoberegger. 2014. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems.* 33–36.

[34] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval.* 83–92.

Paper 2

The following paper is reused from:

- Hernandez-Bocanegra, D.C., & Ziegler, J. (2020). Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics. *Journal of Interactive Media, i-com*, 19(3), 181–200. doi: https://doi.org/10.1515/icom-2020-0021

Research Article

Diana C. Hernandez-Bocanegra* and Jürgen Ziegler

Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics

https://doi.org/10.1515/icom-2020-0021

Abstract: Providing explanations based on user reviews in recommender systems (RS) may increase users' perception of transparency or effectiveness. However, little is known about how these explanations should be presented to users, or which types of user interface components should be included in explanations, in order to increase both their comprehensibility and acceptance. To investigate such matters, we conducted two experiments and evaluated the differences in users' perception when providing information about their own profiles, in addition to a summarized view on the opinions of other customers about the recommended hotel. Additionally, we also aimed to test the effect of different display styles (bar chart and table) on the perception of review-based explanations for recommended hotels, as well as how useful users find different explanatory interface components. Our results suggest that the perception of an RS and its explanations given profile transparency and different presentation styles, may vary depending on individual differences on user characteristics, such as decision-making styles, social awareness, or visualization familiarity.

Keywords: Recommender systems, user study, explanations

1 Introduction

Providing explanations of the rationale behind a recommendation can bring several benefits to recommender systems (RS), by increasing users' perception of transparency (the system explains how it works), effectiveness (user can make good decisions), and trust [38]. Accordingly, research on explainable RS aims to establish methods and models which allow for generating relevant recommendations to users, while providing them with the reasons why an item is recommended. In this regard, various explainable recommendation methods have been proposed, mainly based on collaborative filtering (CF) and content based (CB) methods. CF explanatory models allow to generate explanations based on relevant users or items, e. g. nearest-neighbor style explanations as proposed by Herlocker et al. [21], while CB models facilitate the generation of feature-based explanations by providing users with product features that match their preferences, as proposed, for example, by Vig et al. [44]. On the other hand, matrix factorization (MF) methods, a particular case of CF, allow to generate recommendations by obtaining latent representations of items and users (latent features), which represent a major challenge when it comes to explaining to users how the algorithm works, or why the item is recommended, compared to more intuitive neighbor-style CF or CB methods. In this respect, MF explanatory methods have been proposed [53, 46], to integrate external sources of information (e. g. user generated reviews) in order to make sense – to some extent – of latent features, for example, by aligning them with explicit features drawn from reviews. In this regard, the interest in the use of user reviews in explanation methods has increased recently, given the richness of information reported on diverse aspects, which cannot be deduced from the overall item ratings, and that could be beneficial to both recommendation and explanation processes. Particularly, review-based explanatory approaches usually involve the detection and aggregation of both positive and negative opinions regarding different aspects or features of items, the selection of helpful reviews or excerpts from them that could work as explanations, or the generation of verbal summaries of items' evaluation by users. The above entails a potential for the generation of a diverse range of explanation types, consisting of arguments with different levels of detail, and portrayed in different presentation styles. Nevertheless, little is known

***Corresponding author: Diana C. Hernandez-Bocanegra,** University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany, e-mail: diana.hernandez-bocanegra@uni-due.de, ORCID: https://orcid.org/0000-0002-1773-2633
Jürgen Ziegler, University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany, e-mail: juergen.ziegler@@uni-due.de

about how best to convey explanatory information, in order to meet different explanatory aims like transparency, effectiveness, satisfaction or trust. This is largely due to the predominant lack of evaluation by users in works that propose new explanation methods, as noted by Nunes and Jannach [30]. In this regard, evaluating explanations from the users' perspective can contribute to better explanation design, which can significantly impact users' perception of a RS. Such perspective could contribute to answering questions that remain open, for example, to what extent the format or presentation style influences the perception of an explanation, or what are the components of an explanation that most contribute to its perceived usefulness.

As outlined by Nunes and Jannach [30], explanations may involve the following types of user interface components: input parameters, knowledge base (background or user knowledge), decision inference process (data or rationale of the inference method), and decision output. As for the knowledge base components, and depending on the method used to generate the recommendations, the explanations may reflect either the quality and the properties of the items or the matching between the recommended items and the users' preferences. In regard to the latter, a target user might benefit from knowing which of her/his performed interactions with the system have an effect on a current recommendation [3], as well as knowing how well their preferences match the justifications provided by the system, which can contribute to the acceptance of its recommendations (provided that there is actually a fit) [19]. Although the effects of providing a view on user profiles in CB or item-based CF methods has been explored before (e. g. [3, 39, 18]), such effects have not been fully addressed in review-based explanations, where information on users' profile is often omitted and used only implicitly, e. g. to filter and sort lists of relevant features, as in [28]. In consequence, we aim to address in this article the following question:

RQ1: How does including the information about *user preferences* influence the perception of a review-based RS and its explanations?

Specifically, information on user preferences refers – in the scope of this article – to a list of the relevant inferred aspects and their relevance score, which are also calculated based on the users' own reviews. Additionally, the above mentioned perception is addressed in this article in regard to explanation quality, and to the perception of the overall system in terms of: transparency, effectiveness, efficiency and trust. Likewise, we address the perception of users in regard to specific aspects of the explanations, i. e.: confidence, transparency, satisfaction, persuasiveness, effectiveness, efficiency and easiness to understand of the explanations.

In regard to the interface component "decision inference process" [30], the RS may provide details on the recommendation process, or on the data used for it. In the former case, for example, CF methods favore the generation of concise reporting of recommendation process e. g. "We suggest this option because similar users liked it.". While further algorithm details are often omitted, providing only information about the input and the output of the process might also be beneficial to users, in the case of black-box models [21]. In consequence, various explanatory methods provide information on the data used during the process, like ratings for similar items or ratings by similar users in CF models, or specifications of items in CB methods. However, when additional sources of information are taken into account, as in the case of review-based methods, users are often not informed of the type of the data utilized during the process, for example, whether the user's preferences have been calculated exclusively based on ratings, with information extracted from reviews, or based on other previous interaction with the system. Consequently, we aimed to test to what extent providing such information explicitly is considered useful by the users, more formally:

RQ2: How useful is it for users, during their evaluation of different purchase or booking options, to be informed about the origin of the data used by a review-based recommendation process?

In particular, within the scope of this article, we address how useful it is for users to read that the recommendation is based on the opinions of other customers, as well as their own comments.

The taxonomy of explanations proposed by Nunes and Jannach [30] also involves a category for presentation format, which includes: natural language (e. g. canned text, template-based, structured language), visualization, or other media formats, such as audio. While some of the existing review-based explanatory methods apply at least one of such formats, a user-centered evaluation in which the different formats are comprehensively compared is still necessary. For example, it is not yet clear whether users have a better perception of explanations consisting of aggregate information represented in tabular data, compared to those containing a graphical representation of such information. In this regard, according to Blair [4], visual arguments – defined as a combination of visual and verbal communication – may, in addition to representing propositional content, have a greater "rhetorical power potential" than verbal arguments, due (among others) to their greater immediacy. However, users with lower visual abilities might benefit less from a presentation based on images or graphics [34, 23]. Additionally, while a representation using tables has been recommended to display small data sets [16, 43], if providing accurate numerical values of proportions is not the main objective, tables seem to be less useful than graphics as a means of displaying information [36]. Nevertheless, although the findings in such direction in the field of information visualization, little is known about such effects in relation to explanations. Consequently, we aim to address in this article the following question:

RQ3: How does the display *style* of explanation (using a table or a bar chart) influence the perception of the variables of interest?

Here, the perception of the variables of interest refers to the perception of the overall system and of the specific aspects of explanations, in the same way as described for RQ1.

As it has been shown that individual user characteristics can lead to different perceptions of a RS [25, 50], we assumed that this would also be the case for explanations, as discussed by [2, 26, 22]. Consequently, and similar to Hernandez-Bocanegra et al. [22], we also aimed to test the effect that user characteristics may have on the perception of the explanations, in particular regarding decision making style (rational and intuitive) [20] and the ability of the user to take into account the views of others (social awareness) [17]. Additionally, we also aimed to test the influence that visual familiarization may have on explanations perception, as addressed by Kouki et al. [26]. Consequently:

RQ4: Do individual differences in visual familiarity, social awareness or decision making styles influence the perception of our proposed explanations design?

Here, as with previous RQs, the perception of our explanations designed is addressed in terms of system perception as well as of perception of specific aspects of explanations.

In order to address these questions, we conducted a user study to test the perception of explanations based on user opinions in the hotel domain, given different display styles and whether or not user profile information is shown. The perception was assessed regarding two levels: 1) overall system and explanation quality, and 2) perception of specific aspects of explanations.

The contributions of this paper can be summarized as follows:

- We evaluated the effect of different presentation styles, namely tabulated data or bar charts. Comparisons were conducted both between groups and within participants.
- We also evaluated the effect of providing user profile information as part of explanations, with a display that contains no information regarding user preferences.
- Furthermore, we analyzed the usefulness perceived by users of the different user interface components included in explanations.

The remainder of this paper continues as follows: We discuss related work in Section 2, and the specifics of our explanation design in Section 3. In Section 4, we present methods and results of experiment 1, while details and results of experiment 2 are provided in Section 5. Discussion of both studies and limitations are included in Section 6. Finally, we address future work in Section 7.

2 Related Work

Traditionally, many approaches to explaining the products or services suggested by an RS have been based on ratings provided by users (CF methods) or properties of the recommended items (CB methods). In the former, explanations are often provided in a nearest-neighbor style (e. g. “Your neighbors’ ratings for this movie” [21]), while the latter approach enables the generation of feature-based explanations, that inform users about item properties that may match user preferences, as in [44]. On the other hand, there has recently been increased interest in exploiting alternative sources of information to improve the performance and explainability of RS, particularly the use of user reviews, given the wealth of detailed reports on the positive and negative aspects of an item, information that is often difficult to understand from the general ratings given by users.

Review-based methods enable the generation of the following types of explanations:

1) A verbal summarization of review findings, i. e. statements generated in natural language representing a summarized version of the original content extracted from reviews, e. g. [6, 12], who proposed methods based on natural language generation (NLG) techniques.

2) A selection of helpful reviews, or excerpts from them, that might be relevant to users, as proposed by [9], who used a deep learning model and word embeddings to jointly learn user preferences and item properties, and an attention mechanism to detect features that are of most interest to the target user.

3) A summarized view of pros and cons on specific item aspects reported by other users. Here, topic modelling and aspect-based sentiment analysis are usually used to detect the sentiment polarity towards item aspects or features addressed in reviews, as in [49, 53, 14]. Subsequently, such information can be integrated into RS algorithms such as matrix or tensor factorization, as in [53, 1, 46] in order to generate both recommendation and aspect-based explanations.

In particular, our explanation design proposal and subsequent user study is within the third approach, and is particularly related to the MF model proposed by Zhang et al. [53], since it facilitates the consolidation of statistical information on users' opinions (which can be provided using different presentation styles), as well as their alignment with the user's profile, which is fundamental to our research questions. This model allows the generation of both recommendations and explanations, based on the alignment of 1) latent representations of items and user preferences, and 2) explicit features obtained from reviews. Here, in addition to the rating matrix used in traditional MF, two additional matrices are calculated: a user preference matrix (which indicates how many times a user addressed a feature in their reviews), and an item quality matrix (which indicates how many positive and negative comments were reported in relation to an item). This input information is then used as the basis for our proposed explanation and subsequent user study.

2.1 User Profile Transparency

In regard to providing information on user profile as part of RS explanations, Bilgic and Mooney [3] proposed and tested an influence-based style for explanations in the movies domain, in which the system presents items that had the greatest impact on the recommendation, as well as the ratings that the user has given to those items. They found that such explanations enabled participants to more accurately predict user's satisfaction with the item, compared to a histogram of the user's neighbors' ratings, an explanation style that was found by Herlocker et al. [21] as the best performing among a group of explanations for CF methods, in terms of how compelling they were to study participants.

On the other hand, and using a CB method, Tintarev and Masthoff [39] compared non-personalized verbal explanations with personalized ones, in which, in addition to providing information about the properties of the articles, a sentence was included indicating how these properties related to the user's preferences. According to their findings, personalized explanations were not regarded as more effective than their counter non-personalized part. Here, and similar to [3] the effectiveness was measured based on the difference between the rating that the user would give to an item after reading the explanation, and the one given once the item has been tried. According to authors, the detrimental effect of personalized explanations on effectiveness might be due to users' expectations of preference fit that were not fulfilled once the item was tried.

Additionally, Gedikli et al. [18] compared the perception of users regarding different types of explanations provided by CF and CB recommenders in the movies domain. Their proposed personalized explanations based on clouds showed tags in different colors, depending on the sentiment previously expressed by the target user, regarding different colors (positive: blue, negative: red, neutral: gray). In line with Tintarev and Masthoff [39], they found that a non-personalized tag cloud (all tags in the same color) was slightly more effective than the personalized tag cloud. However, the personalized tag cloud was perceived better by users in terms of transparency.

Disclosure of information used during the recommendation process (e. g. user profile) as part of explanations may facilitate users in identifying and correcting erroneous inferences made by a RS [38]. In this direction, proposed work on scrutable RS seeks to enable and to leverage user control on users' own profile, which in turn may facilitate the generation of new and more accurate recommendations that fit better the real preferences of users. For example, Wasinger et al. [47] implemented a system to recommend restaurant meals based on a scrutable user model, where users could check and adjust their preferences regarding food ingredients to improve recommendations. A user study was conducted to test the application, and noted that users found it easy to understand why certain foods were recommended, by using the customization feature to adjust their preferences.

In regard to review-based methods, Chen and Wang [10] proposed a text-based explanation design that combines both summarization of item opinions as well as item specifications, and that provides a tradeoff view of properties, that allows the direct comparison of different recommended items. They found that a mixed explanatory view containing opinions and specifications was perceived more positively by users, than explanations consisting of only one of such components at the time. However, in contrast to our approach, the selected specifications correspond to explicit elicited preferences, and not to preferences detected from previous reviews written by the user.

On the other hand, Muhammad et al. [28] tested the users' perception of a series of review-based RS explanations in the hotels review. Here, item quality and user preferences are both extracted from reviews and used to generate both recommendations and explanations. However, user preferences are only used implicitly to select, show and sort a subset of the features in explanations, without any mention of such details to users.

In summary, while the effect of presenting explicit information on user profiles as part of explanations of CF and CB methods has already been addressed to some extent, the questions of how such information influences the perception of review-based SR and how such information should be presented remain open.

2.2 Decision-Making Process Transparency

In regard to informing users about the decision inference process, the RS may provide details on the recommendation process, or on the data used for it. Accordingly, Herlocker et al. [21] proposed an explanatory model based on the user's conceptual model of the recommendation process. In a white box conceptual model, users are provided with details of the different steps of the conceptual model of the system operation, e. g. user enters ratings, then system locates similar users, then neighbors' ratings are combined to provide recommendations. In a black box model, however, it may not be practical or even possible to convey details regarding the conceptual model of the system to users, which is actually the case of MF models and their latent features. Herlocker et al. [21] argues that any white box could be regarded as a black box if only information about the input and the output is provided, which could also be beneficial for users.

In regard to the source and type of input used in the process, the presentation of such elements in many of the CF and CB neighbor-style approaches is simpler and self-explanatory, compared to more complex approaches that integrate alternative sources (e. g. reviews) to latent features models as MF, where not only the steps of the process are hard to convey to users, but also the nature of the data used as input. Consequently, most current review-based explanatory approaches omit any mention of the origin of the data, particularly when explaining the inferred user profile, which may make it more difficult to understand compared to item-based explanatory information. Therefore, in addition to assessing how the different ways of presenting the input data might influence the users' perception, in this article we intend to examine also the potential usefulness of explanatory statements on the data origin, as part of review-based explanations, e. g. "based on how often you mentioned features in your own comments before".

2.3 Presentation Format

According to the taxonomy of explanations proposed by Nunes and Jannach [30], explanations could be classified by their presentation format as: natural language (e. g. canned text, template-based, structured language), visualization, or other media formats, such as audio. Regarding review-based explanations, Zhang et al. [53] proposed brief template-based statements to provide information on relevant features (e. g. "You might be interested in [feature], on which this product performs well", although the underlying method allows to generate more detailed explanations, that could also be provided visually using graphs, as elaborated in further sections of the present work. Furthermore, a distinction can also be made between verbal explanations that also provide numerical or statistical information and those that comprise strictly verbal statements. In this respect, Hernandez-Bocanegra et al. [22] compared different types of verbal explanations in the hotel domain, and found that users perceived a higher explanation quality when an aggregated view of positive and negative opinions using percentages was provided, compared with a verbal summary of the opinions that did not provide any percentage, inspired by the abstractive summarization proposed by Costa et al. [12]; furthermore, a greater perceived transparency was reported for explanations with the aggregated view using percentages of opinions, compared to explanations that only provided a useful review, as proposed by Chen et al. [9].

In regard to presentation styles based on visualization techniques applied to review-based RS, Muhammad et al. [28] proposed a summary of the positive and negative opinions on different aspects using bar charts, while Wu and Ester [49] proposed to depict such type of information as word clouds or radar charts. Although bar charts reflecting positive and negative views might be perceived as more informative and attractive than brief template-based textual explanations, easier to interpret than challenging radar charts, or quicker to process than tabulated data, it remains unclear to what extent the presentation format influences the perception of RS and its explanations. In this regard, Kouki et al. [26] proposed a series of explanations based on a hybrid RS in the music domain, and tested, among others, the influence that the presentation format could have on users' perception. In this case, the authors found that textual explanations were perceived as more persuasive than the explanations provided using a visual format; however, users with greater visual familiarity perceived one of the visual format explanations more positively (a Venn diagram). Consequently, we aimed to investigate whether such an effect is also observed in the case of review-based explanations.

Particularly, in the present work, we set our focus on two formats: bar chart and table. Bar charts are recommended to facilitate a direct and quick comparison of values between different categories or items, contrary to alternatives like pie charts, or bubbles, where additional cognitive efforts would be needed to accurately calculate the differences in values across categories, in our case, the different aspects of the items. Likewise, word clouds imply a presentation challenge, since we are willing not only to represent the amount of comments (which could be reflected by font size), but also polarity, which would require using separate clouds for the positive and negative aspects, or showing a single predominant sentiment per aspect in a single cloud, thus obscuring the information about the less predominant polarity. On the other hand, while the use of tables has been recommended to display small data sets (less than 20 data points) [16, 43], when providing exact numbers or proportions is not the main objective, tables seem to be less useful than graphics [36]. As indicated previously in the case of verbal explanations, users benefited from a view that provides percentages of positive and negative opinions, suggesting that percentages may serve as anchors to convey more compelling information in explanations, compared to purely verbal statements. In this sense, when motivation or ability is lacking, the effortless use of cues such as numerical anchors can lead to changes in attitude [32, 48], which in turn influence judgments and decision making (anchoring effect). Thus, even when the values of the proportions of the opinions included in the two types of explanations (table or chart) are the same, a different representation of them might lead to differences in explanation perception, which we set out to test in the user study.

2.4 User Characteristics

Beyond the explanations' content itself, a number of user characteristics also contribute to differences in the overall perception of RS. Models proposed by Knijnenburg et al. [25] and Xiao and Benbasat [50] argue that perception of the interaction with the system usually depends on personal characteristics, like demographics and domain knowledge. Furthermore, Berkovsky et al. [2] evaluated how differences in the perception of trust might reflect differences in users' personality traits, given different types of explanations provided in the movies domain. To this end, they used participants' scores of the Big-Five personality traits (openness, conscientiousness, extraversion, agreeableness and neuroticism) [13, 40], and compared persuasive explanations (e. g. "highest grossing movie of all times"), personalized CF-based explanations (e. g. "because you liked X") and IMDb voting-based explanations (e. g. "Average rating *n*, Number of votes *m*"). Among their findings, authors reported that people with higher disposition to agree perceived more positively the voted-based explanations, compared to personalized explanations, seemingly to a higher disposition to accept others' opinions rather than impose their own preferences. Furthermore, they found that people with higher levels of neuroticism perceived better the voted-based explanations compared to the persuasive ones, possibly due to a perception of higher reliability of explicit voting numbers, which could presumably reduce the risk of frustration of a person with high levels of neuroticism.

Similarly, Kouki et al. [26] explored the influence of personality traits on users' explanation preferences regarding perceived accuracy and perceived novelty of recommendations, in the music domain. They compared different types of textual explanations, and found that participants with higher levels of neuroticism preferred item-based explanations (e. g. "people who listen to your profile item X also listen to Y") whereas popularity-based explanations (e. g. "X is a very popular in the last.fm database with *n* million listeners and *m* million playcounts") were preferred by users with lower levels of neuroticism, the latter in contrast to the opposite finding reported by [2], regarding trust perception.

Despite the usefulness of using the Big Five personality traits to better understand individual differences in RS perception and its explanations, we decided to address other types of user characteristics, which are more related to how users process information when making decisions, noting that supporting this process is precisely the goal of recommendation systems. Particularly, individual differences in decision-making styles are determined to a greater extent by preferences and abilities to process available information [15]. Two main aspects provide a basis to describe the differences in decision styles: information use (amount of information used during the process) and focus (alternatives addressed) [15]. "Good enough" information might be sufficient for some people, whereas others prefer to obtain and address all relevant information, in order to minimize risks or negative consequences of decisions. To the former, even when more information may be available, it is not necessary or worth taking the time to review it. Hamilton et al. [20] defines rational and intuitive decision styles similarly to the cognitive styles of Pacini and Epstein [31], with the latter having a more general scope to describe manners of solving problems. Thus, decision making styles are defined by Hamilton et al. [20] as a "habit-based propensity" to exhaustively search for information and to systematically evaluate possible alternatives (rational style), or to use of a quick process based on hunches and feelings (intuitive style).

Additionally, we were interested in another factor that may influence the way users perceive explanations: the extent to which they are able to adopt the perspective of others when making decisions. The rationale for this interest stems from the tendency of individuals to adjust their own opinions using those of others, while choosing between various alternatives [35], which may even be beneficial [51]. Particularly, individuals with greater perspective-taking skills tend to understand the views of others better [8, 5], skills that are also characterized as "social awareness" [17].

Previous work by Hernandez-Bocanegra et al. [22] evaluated the influence of decision-making styles and social awareness on the perception of review-based argumentative explanations, and suggested that social awareness might have an effect on both transparency and trust in review-based RS. Their results indicated that users with a greater willingness to listen and take into account the opinions of others valued their proposed explanations better than users who tend to listen less to others. On the other hand, contrary to the authors' expectations, the more detailed explanations summarized by the system were not preferred by the more rational users, apparently because the additional information generated by the system is not perceived as more satisfactory than the possibility of reading directly the comments written by the users.

Finally, since we aimed to compare differences in perception of explanations consisting of different visual representations, we also considered a factor that is related to visual abilities, in particular the extent to which a user is familiar with graphical or tabular representations of information. Visualization familiarity may also influence the perception of explanations provided using images or graphs, as found by Kouki et al. [26] in the music domain. Here, authors found that textual explanations were perceived as more persuasive than the explanations provided using a visual format; however, users with greater visual familiarity perceived one of the visual format explanations more positively (in particular a Venn diagram).

3 Explanation Design

In the context of RS, review-based argumentative explanations could be understood as a set of propositions, summarizing positive points reported by other users on specific aspects, that support the claim that an article can be recommended to a user. In this respect, information extracted from user reviews could be consolidated and provided as propositions, which would constitute the *backing* component according to the argumentative scheme proposed by Toulmin [42], while the conclusion (the item is recommended) constitute itself a *claim*. While this could be considered a 'shallow' structure, compared to the complete Toulmin argument scheme (which involves additional components, like rebuttal or refutation), it resembles explanation schemes based on deductive arguments, such as those widely used in the scientific field (i. e. a set of explanatory propositions is logically followed by an explanatory target, as discussed by Thagard and Litt [37]), or even more particularly, explanation schemes in RS such as the one used by Zanker and Schoberegger [52], who provides brief sentences – two facts and a claim – as explanations for content-based recommendations of hiking routes, energy and mobile phone plans.

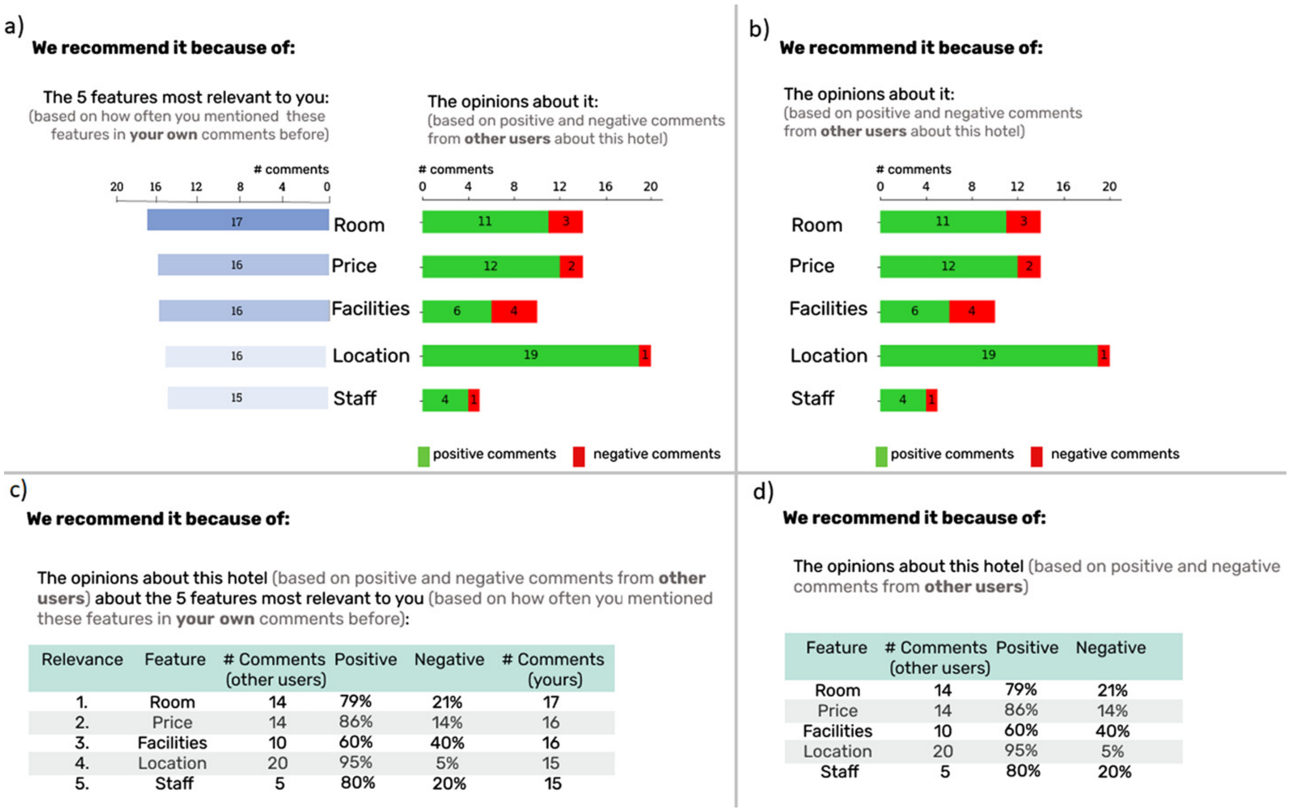
In consequence, our explanation design (see Figure 1) seeks to represent an argumentative structure, while reflecting in turn the arguments provided by other users in their reviews, in a consolidated manner. Therefore, our proposed scheme consists of a claim ("We recommend this hotel") and the propositions that support such claim, connected with the conjunction "because". We propose to provide the following pieces of information in proposition statements:

1. Item quality: A summary of comments reported by previous hotel guests for different aspects, as well as what percentage were positive and negative.

2. User preferences: what are the most important item aspects to the target user. In this regard, we aimed to make the user's own profile transparent, by showing the user's inferred importance of each aspect, together with the opinions of other users about the aspect (as shown in the examples included in Figures 1a and 1c), in order to facilitate a direct comparison of the points of view of others and their alignment with their own preferences.

3. Statements that inform how the user preferences and item quality are extracted (e.g,"based on how often you mentioned these features in your own comments before"). We believe that providing this information, in addition to the information listed above, could increase the perception of trust by users, while decreasing the perception that they are interacting with a black box.

While arguments are usually associated with oral or written speech, arguments can also be communicated using visual representations (e. g. graphics or images). In this


a)
We recommend it because of:
The 5 features most relevant to you:
(based on how often you mentioned these features in your own comments before)
# comments
20 16 12 8 4 0
17
16
16
16
15
Room
Price
Facilities
Location
Staff
The opinions about it:
(based on positive and negative comments from other users about this hotel)
# comments
0 4 8 12 16 20
11 3
12 2
6 4
19 1
4 1
🟩 positive comments 🟥 negative comments
b)
We recommend it because of:
The opinions about it:
(based on positive and negative comments from other users about this hotel)
# comments
0 4 8 12 16 20
Room
Price
Facilities
Location
Staff
11 3
12 2
6 4
19 1
4 1
🟩 positive comments 🟥 negative comments

c)

We recommend it because of:

The opinions about this hotel (based on positive and negative comments from other users) about the 5 features most relevant to you (based on how often you mentioned these features in your own comments before):

Relevance	Feature	# Comments (other users)	Positive	Negative	# Comments (yours)
1.	Room	14	79%	21%	17
2.	Price	14	86%	14%	16
3.	Facilities	10	60%	40%	16
4.	Location	20	95%	5%	15
5.	Staff	5	80%	20%	15

d)

We recommend it because of:

The opinions about this hotel (based on positive and negative comments from other users)

Feature	# Comments (other users)	Positive	Negative
Room	14	79%	21%
Price	14	86%	14%
Facilities	10	60%	40%
Location	20	95%	5%
Staff	5	80%	20%

Figure 1: Explanations displayed in empirical study for every experimental condition, for one of the recommended hotels. a) Style 'visual', user preferences 'yes'. b) Style 'visual', user preferences 'no'. c) Style 'text', user preferences 'yes'. d) Style 'text', user preferences 'no'.

regard, according to Blair [4], visual arguments (a combination of visual and verbal communication) may, in addition to representing propositional content, have a greater "rhetorical power potential" than verbal arguments, due (among others) to their greater immediacy.

In consequence, we aimed to test the effect of the two factors: display *style* and display of the *user preferences*. An example of each condition is provided in Figure 1.

'Bar chart' style: Provides a view of the number of comments per aspect and percentages of positive and negative opinions using bar charts.

'Table' style: Provides the same information used in the visual condition, but instead of using bar charts, presents the information within a table.

Additionally, every display style involves two variations:

User preferences 'yes'. The information about the user preferences is provided.

User preferences 'no'. No information about the user preferences is displayed.

4 Experiment 1: System and Explanation Quality Perception, between Subjects

We implemented a prototype of a hotel recommender system that provides both recommendations and explanations, based on the design discussed in Section 3, and conducted an experiment where we compared users' perception of the overall system in terms of transparency, effectiveness, efficiency and trust. In this regard, we aimed to test our hypothesis that users would report a more positive perception of the RS when information about their user preferences is provided (*H1*). Additionally, we hypothesized that users with greater visual abilities would find explanations better when these are provided using visual aids, such as a bar chart, in comparison to tabulated information (*H2*). In particular, the aim of experiment 1 was to compare the overall perception of the prototype and its explanation quality, in a between groups manner (participants were assigned to conditions that reflect the different types of explanations designed), while in a subsequent experiment (see Section 5) we addressed the perception of specific aspects of explanations within subjects, as well as the usefulness of individual explanation components.

4.1 Methods

Participants

We recruited 150 participants (66 female, mean age 39.08 and range between 23 and 73) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 95 %, and number of HIT's approved greater than 500. We applied a quality check in order to select participants with quality survey responses, i. e. at least 5 of the 6 high priority validation questions were answered correctly, more than 30s were spent on the recommendation step and more than 50s on the evaluation questionnaire. The responses of 46 subjects were discarded due to this quality check (from an initial number of 195 workers), so only the responses of 150 subjects were used for the analysis (statistical power of 85 %, $\alpha = 0.05$). Participants were rewarded with $1 plus a bonus up to $0.4 depending on the quality of their response to the question "Why did you choose this hotel?" set at the end of the survey. Time devoted to the task by participants (in minutes): M = 8.04, SD = 1.62.

Study Design

The study follows a 2x2 between-subjects design, and each participant was assigned randomly to one of four conditions that represent the combination of the two factors: display *style* and *user preferences* provided or not. We presented participants with a fixed list of 5 hotels that represented the recommendations for a hypothetical hotel search, and a detailed view including an explanation of why every item was recommended. Then, participants were asked to choose the hotel they considered the best, to report their reasons to it, and to rate their perception of both recommender and its explanations. The explanations and recommendations were generated using the EFM algorithm [53] and the dataset of hotels' reviews, ArguAna [45], although they were presented to the participants only through a prototype, i. e. no real system was implemented to allow the interactions.

Given that we had no access to previously written participants' reviews (which is not only important for the optimal functioning of the algorithm, but also constitutes a base to test the condition "user preferences"), we calculated the top 5 of the most important aspects to all users within the dataset, namely: room, price, facilities, location and staff. Then, a random user was chosen from the dataset with those same preferences, and 5 of her top-ranked options according to the EFM algorithm were selected to be presented to participants, alongside their explanations. Additionally, we presented the users with a cover story, in which we told the users to pretend that their most important aspect was the "room" and the "price".

Questionnaires

Evaluation: We utilized items from [33] to evaluate the perception of system transparency (construct *transparency*, user understands why items were recommended), from [25] to evaluate the perception of system effectiveness (construct *perceived system effectiveness*, system is useful and helps the user to make better choices) and efficiency (user can save time with the recommender), and items from [27] to assess the perception of trust in the system (constructs *trusting beliefs*, user considers the system to be honest and trusts its recommendations, and *trusting intentions*, user willing to share information). In addition, we also adapted 3 items from [25] to address explanation quality (construct *perceived recommendation quality*, user likes explanations, considers them relevant). All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

User characteristics: We used all the items of the Rational and Intuitive Decision Styles Scale [20] as well as the scale of the social awareness competency [17]. Additionally, We used the visualization familiarity items as proposed by [26]. All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

Procedure

First, participants were asked to answer demographic questions and the questionnaire on user characteristics. We indicated in the instructions step that a 5 hotels list reflecting the results of a hypothetical hotels' search would be presented. We asked them to click the "View Details" button for each hotel, and to read carefully the explanations provided in each case (examples of explanations for the different experimental conditions are provided in Figure 1). Additionally, we provided a cover story, as an attempt to establish a common starting point in terms of reasons to travel (a business trip), and the supposedly most interesting aspects for the user (room and facilities).

The list of hotels, their names, photos, prices and locations, as well as their ratings and the numbers of reviews and positive and negative opinions, remained constant to all users. Variations focused only on display style and the presentation of user preferences, depending on the condition to which each participant was assigned. After the interaction with the prototype, subjects were asked to choose the hotel that best suited their purpose, as well as an open question about their reasons for choosing that hotel. Then, subjects answered the evaluation questionnaire. In addition, we included an open-ended question, so that participants could indicate in their own words their general opinion about the explanations provided. We included 11 validation questions to check attentiveness and the effective completion of the task.

Data Analysis

We evaluated the effect that display style and the display of user preferences (independent variables IVs) may have on the perception of the prototype and its explanations, and to what extent user characteristics (regarded as moderators or covariates) could influence such perception (rational and intuitive decision-making style, social awareness and visualization familiarity). Here, the dependent variables (DVs) are evaluation scores on: system transparency, effectiveness, efficiency, trust and explanation quality. Here, evaluation scores were calculated as the average of the individual values reported for the questionnaire items related to each DV. Regarding the covariates, we calculated the scores of the rational and the intuitive decision making styles, social awareness and visualization familiarity for each individual as the average of the reported values for the items of every scale.

Given that our DVs are continuous (scores are the averages of reported answers of questionnaire items of each construct) and correlated (correlation coefficients in Table 1), and that we address also the effect of covariates, a MANCOVA analysis was performed, to assess the simultaneous effect of presentation styles and interactivity on the overall system perception, as well as the influence of user characteristics on it. Subsequent ANCOVA analyses were performed to test main effects of IVs and covariates, as well as the effect of interactions between them. Q-Q plots of residuals were checked to validate the adequacy of the analysis.

4.2 Results

Evaluation and User Characteristics Scores

We found that explanations including information of user preferences are perceived slightly better than explanations without this information in terms of explanation quality and system transparency, effectiveness, efficiency and trust. On the other hand, explanations including a bar chart were perceived slightly better than explanations with a table, in regard to explanation quality and trust, while the opposite was observed in relation to transparency, effectiveness and efficiency. However, as discussed in detail below, such differences are not statistically significant. The average evaluation scores by presentation style and display of user preferences are shown in Table 1.

In regard to user characteristics, distributions of the scores of rational (M = 4.24, SD = 0.56) and the intuitive (M = 2.65, SD = 1.01) decision making styles, social awareness (M = 3.92, SD = 0.59) and visual familiarity (M = 3.03, SD = 1.02) are depicted in Figure 2a. Here, we observed a skewed right distribution of rational decision making-style and social awareness scores, not being that the case for the intuitive decision-making style and visual familiarity, i. e., most users consider themselves to be predominantly rational decision makers who are able to listen to others and take into account the opinions of others; however, a more balanced distribution is observed in the remaining user characteristics: only a minority recognize themselves as very (or not at all) familiar with visual representations of information, and as very (or not at all) intuitive decision makers. In addition, results suggest an influence of some of the user characteristics on the perception of the system by users, which we describe in detail below.

Table 1: Experiment 1, mean values and standard deviations of perception on explanation aims, per *display style* and *display of user preferences* (n = 150); values reported with a 5-Likert scale; high mean values correspond to a positive perception of recommender and its explanations. Pearson correlation matrix, p<0.001 for all correlation coefficients.

Style	Table		Bar chart		*User Pref.*	Yes		No		*Corr.*	Variable			
Variable	M	SD	M	SD		M	SD	M	SD		1	2	3	4
1. Expl. Quality	3.83	0.65	3.86	0.67		3.88	0.67	3.81	0.65					
2. Transparency	3.96	0.73	3.87	0.85		3.99	0.74	3.84	0.84		0.37	—		
3. Effectiveness	3.88	0.61	3.75	0.75		3.84	0.76	3.79	0.61		0.60	0.47	—	
4. Efficiency	3.96	0.73	3.89	0.92		4.00	0.78	3.86	0.87		0.36	0.39	0.52	—
5. Trust	3.75	0.60	3.89	0.63		3.84	0.65	3.81	0.59		0.66	0.40	0.67	0.58

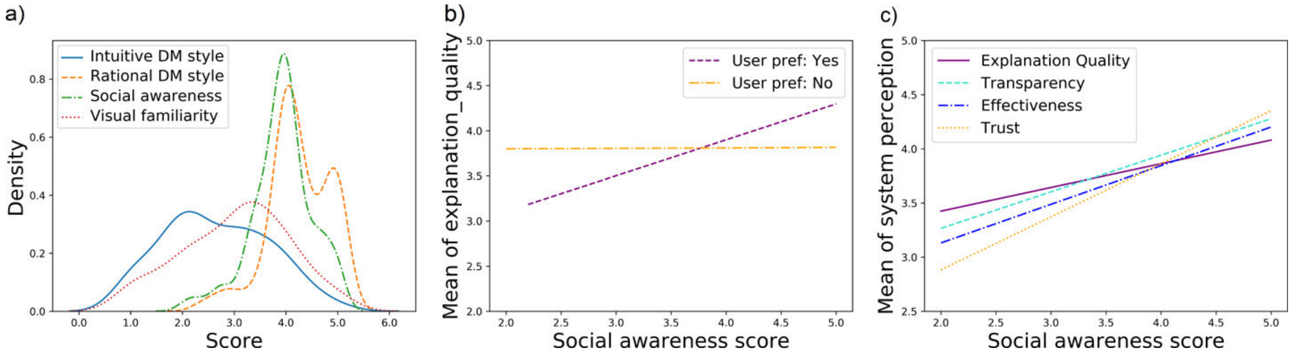


Figure 2: Plots Experiment 1. a) Kernel density estimate of user characteristics scores: rational and intuitive decision making styles, social awareness and visual familiarity. b) Interaction plot for explanation quality (fitted means of individual scores) between display of user preferences and social awareness. c) Effect of social awareness on all explanation aims (fitted means of individual scores). All scores within the range [1,5].

System and Explanation Quality Perception

Presentation style and display of user preferences: We found no main significant effect of the combination of these factors.

Display of user preferences: No significant multivariate effect was found for display of user preferences.

Presentation style: No significant multivariate effect was found for presentation style.

Rational decision-making style: We found a significant main effect of rational style $F(5, 138) = 4.50$, $p < .001$. Univariate tests revealed a significant effect of this variable on: effectiveness $F(1, 142) = 9.12$, $p = .003$), efficiency ($F(1, 142) = 10.98$, $p = .001$) and trust ($F(1, 142) = 18.82$, $p < .001$). Here, a positive trend was observed between rational decision-making score and the above mentioned DVs, i. e. the higher the rational decision making score, the higher the perception scores of these DVs.

Intuitive decision-making style: We found a significant main effect of intuitive style $F(5, 138) = 3.25$, $p = .008$. Univariate tests revealed a significant effect of this variable on: explanation quality ($F(1, 142) = 16.37$, $p < .001$). Here, a positive trend was observed between this variable and the score of intuitive decision-making style.

Social awareness: We found a significant main effect of social awareness $F(5, 138) = 6.72$, $p <.001$. Univariate tests revealed a significant effect of this variable on: explanation quality ($F(1, 142) = 5.62$, $p = .019$), transparency ($F(1, 142) = 7.93$, $p = .006$), effectiveness ($F(1, 142) = 8.79$, $p = .004$) and trust ($F(1, 142) = 26.56$, $p < .001$). Here, we observed a positive trend in the relationship between social awareness and these DVs (see Figure 2d).

Additionally, a significant interaction effect between social awareness and the display of user preferences on explanation quality was found $F(1, 146) = 4.79$, $p = .030$, with the "yes" condition having a steeper slope than the "no" condition (showing a positive relationship between social awareness and displaying user preferences), the latter remaining constant regardless of the social awareness score (Figure 2b).

5 Experiment 2: Perception on Specific Aspects of Explanations, within Subjects

We used screenshots of the prototype implemented for experiment 1 (see Section 4), reflecting the design discussed in Section 3, and conducted a second experiment aiming to compare users' perception of specific aspects of explanations, when presented to all the four possible explanations (see Figure 1). In experiment 2, differences were addressed within subjects, while in experiment 1 we evaluated the perception of the overall system in a between subjects manner. Likewise to experiment 1, we also aimed to test our hypothesis that users would report a more positive perception when information about user preferences is provided (*H1*), and also that users with greater visual abilities would find explanations better when these are provided using visual aids, such as a bar chart, in comparison to tabulated information (*H2*).

Additionally, the experiment 2 also involved the assessment of the usefulness of individual components of explanations, by participants. In this regard, for example, we hypothesised that most users would find useful the information regarding the origin of the explanatory information provided (*H3*).

5.1 Methods

Participants

We recruited 35 participants (14 female, mean age 42.77 and range between 24 and 65) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 95 %, and a number of HIT's approved greater than 500. We applied a quality check in order to select participants with quality survey responses, i. e. at least 5 of the 7 validation questions were answered correctly. The responses of 7 subjects were discarded due to this quality check (from an initial number of 42 workers), so only the responses of 35 subjects were used for the analysis, a value consistent to our within subjects design (statistical power of 95 %, $\alpha = 0.05$). Participants were rewarded with $1. Time devoted to the task by participants (in minutes): M = 6.70, SD = 1.07.

Study Design

The study follows a within-subjects design, and each participant was presented sequentially with an example of each of the 4 types explanations, that represent the combination of the two factors: display *style* and *user preferences* provided or not. The order of presentation of the 4 types of explanation was counterbalanced.

Questionnaires

We used the user experience items (UXP) proposed by [26] to address the explanations reception, comprising: explanation confidence (explanation makes user confident that she/he would like the recommended item), explanation transparency (explanation makes the recommendation process clear), explanation satisfaction (user would enjoy a system if recommendations are presented this way), and explanation persuasiveness (explanations are convincing). Finally, we included additional elements to assess explanation effectiveness (user can make better decisions if explanation presented this way), explanation efficiency (user can save time if system provides this type of explanation), and explanation easiness (explanation is easy to understand). All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree). Users were asked to respond to the same user characteristics questionnaire we used in experiment 1.

Additionally, participants were requested to provide their opinions on how helpful they considered the different components of the explanations: the bar plots, the tables, the information about others' opinions, the information about their supposed own comments, and the information on where the bar plots and tables come from. All items were measured with a 1–5 Likert-scale (information is helpful, 1: Strongly disagree, 5: Strongly agree).

Procedure

First, participants were asked to answer demographic questions and the questionnaire on user characteristics. We indicated in the instructions that they will be presented with information about the pros and cons of different hotel features that might be relevant to you, using 4 different display options, and that they would then indicate their opinion about each option. Additionally, we provided a cover story, as an attempt to establish a common starting point in terms of reasons to travel (a business trip), and the supposedly most interesting aspects for the user (room and facilities). After the assessment of all types of explanations, participants were asked to reply questions about the usefulness of specific components of explanations. At the end, they were asked to report their comments and suggestions about the explanations with an open-ended question.

Data Analysis

We evaluated the effect that display style and the display of user preferences (independent variables IVs) may have on the perception of specific aspects regarding the proposed explanations, and to what extent user characteristics (regarded as moderators or covariates) could influence such perception (rational and intuitive decision-making style, social awareness and visualization familiarity). Here, the dependent variables (DVs) are evaluation scores on the following aspects: explanation confidence, explanation transparency, explanation satisfaction, explanation persuasiveness, explanation effectiveness, explanation efficiency and explanation easiness to understand. Regarding the covariates, we calculated the scores of user characteristicas the same way as in study 1 (average of the reported values for the items of every user characteristics scale).

Given that our DVs are ordinal (scores are the reported answers to single questionnaire items) we performed a Friedman test, the non-parametric alternative to the repeated measures ANOVA. Given that our variables are correlated, the significant tests were conducted using Bonferroni adjusted alpha levels of .007 (.05/7).

Additionally, we calculated the average evaluation scores for each possible value of the two factors: presentation style (bar chart and table), and display of user preferences (yes and no). Using these continuous and correlated evaluation scores, we then perform a repeated measures MANCOVA, to assess the simultaneous effect of presentation styles and display of user preferences on explanations perception, as well as the influence of user characteristics on it. Subsequent ANCOVA analyses were performed to test main effects of IVs and covariates, as well as the effect of interactions between them.

Usefulness of explanations components: We performed a series of ordinal logistic regressions to test influence on scores of usefulness of components – DVs (bar chart, table, others' opinion view, own preferences view, information source) by predictor variables, in this case the user characteristics (rational and intuitive decision-make style, social awareness and visualization familiarity), which were tested a priori to verify there was no violation of the assumption of no multicollinearity. Q-Q plots of residuals were also checked to validate the adequacy of the analysis.

DVs were initially rated using a 5-likert scale, but additionally we grouped answers as Yes (agree and strongly agree that element is helpful), and No / Neutral (disagree, strongly disagree and neutral that element is helpful) for subsequent analysis. We then calculated the percentages of Yes and No/Neutral responses regarding the different explanation components, and performed a binomial test, to check whether the proportions of Yes and No/Neutral answers were different from a proportion that assumes that the percentages are equal (50 % of Yes and 50 % of No/Neutral).

Finally, we used a Wilcoxon rank t-test to compare the average responses of the perception of usefulness of a view of others' opinion with that of a view of their own preferences, as well as the average responses of perceived usefulness of tables compared to bar charts in explanations.

5.2 Results

Evaluation and User Characteristics Scores

We observed only small differences between table and bar chart explanations, and between explanations including or not user preferences, in regard to most of the specific aspects of explanations evaluated, with the exception of easiness to understand. As discussed in detail below, explanations without display of user preferences were perceived easier to understand, this difference being statistically significant. The average evaluation scores by presentation style and display of user preferences are shown in Table 2.

In regard to user characteristics scores, we observed similar distributions of such scores: a skewed right distribution of rational decision making-style and social awareness scores, not being that the case for the intuitive decision-making style and visual familiarity. Distributions of the scores of rational (M = 4.34, SD = 0.7) and intuitive (M = 2.13, SD = 0.83) decision making styles, social awareness (M = 3.55, SD = 0.53) and visualization familiarity (M = 2.82, SD = 1.17) are depicted in Figure 3a. Additionally, we observed a main effect of some of these user characteristics on the perception of specific aspects of explanations, as well as interaction effects involving these variables. Such findings are described below.

Perception of Explanations

Presentation style and display of user preferences: We found no main significant effect of the combination of these factors after Bonferroni correction.

Display of user preferences: We found a multivariate effect of display of user preferences, $F(7,28) = 2.41$, $p = .046$. Univariate tests revealed a main effect of display of user preferences on explanation easiness to understand $F(1,34) = 6.42$, $p = .016$, so that explanations that do not include information on user preferences are significantly easier to understand (M=3.76, SD=0.91), compared to those showing such information (M=4.01, SD=0.72).

Presentation style: No multivariate main effect of presentation style was found.

Table 2: Experiment 2, mean values and standard deviations of perception on explanation aspects, per *display style* and *display of user preferences* (n=35); values reported with a 5-Likert scale; high mean values correspond to a positive perception of explanations aspects.

Style	Table		Bar chart		*User Pref.*	Yes		No	
Variable	M	SD	M	SD		M	SD	M	SD
1. Expl. Confidence	3.30	0.92	3.33	0.98		3.33	0.86	3.30	0.96
2. Expl. Transparency	3.50	0.88	3.64	0.97		3.51	0.88	3.63	0.74
3. Expl. Satisfaction	3.41	1.08	3.27	1.05		3.24	1.03	3.44	0.93
4. Expl. Persuasiveness	3.30	0.92	3.29	1.02		3.34	0.94	3.24	1.06
5. Expl. Effectiveness	3.33	0.97	3.37	0.95		3.29	0.93	3.41	0.85
6. Expl. Efficiency	3.31	1.13	3.34	1.09		3.21	1.05	3.44	0.93
7. Expl. Easiness to understand	3.76	0.92	3.84	0.86		3.59	0.93	4.01	0.72

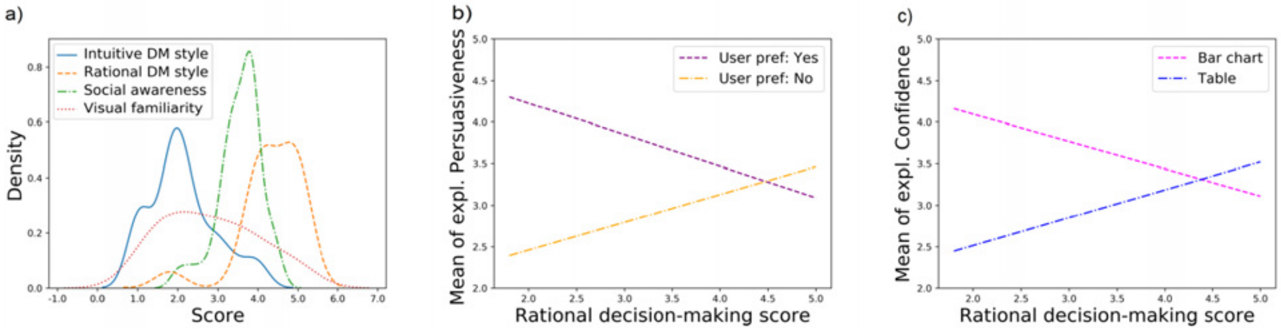


Figure 3: Plots Experiment 2. a) Kernel density estimate of user characteristics scores: rational and intuitive decision making styles, social awareness and visual familiarity. b) Interaction plot for explanation persuasiveness (fitted means of individual scores) between display of user preferences and rational decision-making style. c) Interaction plot for explanation confidence (fitted means of individual scores) between presentation style and rational decision-making style. All scores within the range [1,5].

Display of user preferences and rational decision-making style: We found a multivariate interaction effect between these two variables, $F(7,27)$ =, $p = .002$. Univariate tests revealed the significant interaction effect ot these variables on: explanation transparency ($F(1,33) = 7.79$, $p = .009$), explanation satisfaction ($F(1,33) = 5.62$, $p = .024$), explanation persuasiveness ($F(1,33) = 20.67$, $p < .001$), explanation easiness to understand ($F(1,33) = 7.36$, $p = .011$) and explanation effectiveness ($F(1,33) = 5.34$, $p = .027$). For all these DVs, the same trend was observed: the higher the reported rational decision-making style, the higher the scores on the different DVs when the user profile was not shown, while the opposite trend was observed when it was shown. An example of this trend is observed in Figure 3b.

Presentation style and rational decision-making style: We found a multivariate interaction effect between these two variables, $F(7,27)$ =, $p = .006$. Univariate tests revealed the significant interaction effect ot these variables on: explanation confidence ($F(1,33) = 14.09$, $p = .001$), explanation satisfaction ($F(1,33) = 5.78$, $p = .022$), explanation easiness to understand ($F(1,33) = 7.36$, $p = .011$), explanation effectiveness ($F(1,33) = 4.34$, $p = .045$) and explanation efficiency ($F(1,33) = 7.15$, $p = .012$). For all these DVs, the same trend was observed: the higher the reported rational decision-making style, the higher the scores on the different DVs when the table was provided, while the opposite trend was observed when the bar chart was shown. An example of this trend is observed in Figure 3c.

Usefulness of Explanation Components

Effect of user characteristics on usefulness.

An increase in intuitive decision-making score was significantly associated with an increase in the odds of participants reporting higher values of: usefulness of bar charts in explanations, with an odds ratio of 3.16 (95 % CI, 1.12 to 9.81), Wald $\chi2(1) = 4.76$, $p = .029$, and usefulness of a view of others' opinions in explanations, with an odds ratio of 3.21 (95 % CI, 1.11 to 11.19), Wald $\chi2(1) = 4.69$, $p = .030$.

An increase in social awareness score was significantly associated with an increase in the odds of participants reporting higher values of: usefulness of information origin in explanations, with an odds ratio of 5.77 (95 % CI, 1.38 to 27.20), Wald $\chi2(1) = 5.82$, $p = .016$.

An increase in visualization familiarity score was significantly associated with an increase in the odds of participants reporting higher values of: usefulness of bar charts in explanations, with an odds ratio of 3.79 (95 % CI, 1.76 to 9.47), Wald $\chi2(1) = 12.33$, $p < .001$, usefulness of a view of own comments in explanations, with an odds ratio of 2.54 (95 % CI, 1.35 to 5.14), Wald $\chi2(1) = 8.62$, $p = .003$, and usefulness of information origin in explanations, with an odds ratio of 3.21 (95 % CI, 1.62 to 6.96), Wald $\chi2(1) = 11.77$, $p < .001$.

Participants who found components helpful.

We found then that a significant majority found the information about others' opinion helpful ($\chi2(1, N = 35) = 6.40$, $p = .011$), so that both tables ($\chi2(1, N = 35) = 6.40$, $p = .011$) and bar charts ($\chi2(1, N = 35) = 8.75$, $p = .003$), whereas only a significant minority found the display of user preferences helpful ($\chi2(1, N = 35) = 6.40$, $p = .011$). On the other hand, providing details about where the information used for the recommendation comes seems to be regarded as helpful by most people, but the difference with the proportion of people that found it non helpful / neutral is not significant. Proportions are depicted in Figure 4.

Comparison of usefulness scores.

We found that the average usefulness of the components view of others' opinions and view of own prefer-

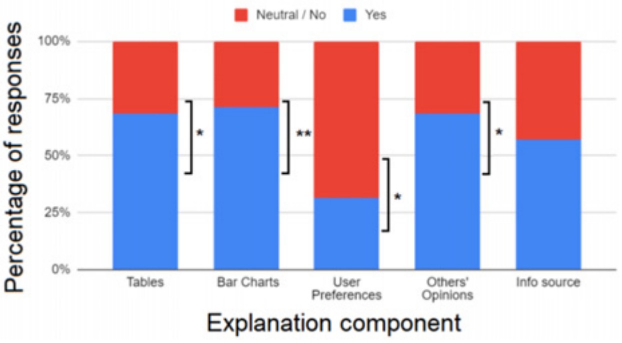


Figure 4: Proportion of participants who found the different explanation components helpful (Y) or non helpful neutral (Neutral/No). * p <0.5, **p<0.01.

ences are significantly different ($W = 0.89$, $p < .001$), with the display of others' opinions having a higher mean (M = 3.74, SD = 1.07) than the display of users' preferences (M = 2.63, SD = 1.29). On the other hand, we found no significant difference when comparing the mean responses of usefulness of tables (M = 3.60, SD = 1.19), and bar charts (M = 3.71, SD = 1.20), although bar charts are perceived slightly more helpful than tables.

6 Discussion

6.1 Effect of Profile Transparency

In regard to our **H1**, we found no main effect of the display of user preferences on the perception of the system or its explanations. Although contrary to our expectations, the lack of a significant influence of disclosing user preferences seems to be somehow in line with the results reported by Tintarev and Masthoff [39], Gedikli et al. [18], who observed that providing personalized explanations (in which preferences were presented along with item properties), while potentially beneficial in terms of satisfaction with the explanations, did not necessarily result in a better perception of effectiveness (helping the user to make better decisions). The authors suggested that a possible reason could be a mismatch between the expectation generated by the explanation and the actual evaluation after trying the item. In our case, however, this could be related to how easy it was for the participants to understand the explanations. In particular, we observed that explanations without information on user preferences were significantly easier to understand compared to those that included such information. In addition, we observed that users with less visualization familiarity reported lower usefulness scores of the user preference section, suggesting that the proposed presentation of this section still needs to be improved to benefit users who do not have sufficient experience with information visualization techniques as well.

Although the display does not have a main effect on the perception of the system and its explanations, we observed a mediating effect of social awareness, such that individual differences in this characteristic were reflected in differences in the perception of the explanation quality. Here, our findings suggest that people who tend to listen more to others tend to perceive better the explanations that include information about their own profile. On the other hand, when user preferences are not displayed, the perception of explanation quality remains pretty much the same, despite the extent of users' social awareness. At this respect, we believe that users with greater abilities to take into account the opinion of others might appreciate the chance to see the alignment of their own preferences with the opinions of others, in an effortless manner, given that a metric of aspect relevance was placed right next to the metric of other users' opinions regarding such aspect.

Additionally, although no significant interaction effect was found between the rational decision-making style and the display of user preferences on the perception of the system in general, we found that this interaction had a significant effect on the perception of most of the specific aspects of the explanations. In this case, users who reported higher scores for rational decision-making style reported less preference for explanations that provided information on the user's profile. In this regard, we believe that more skeptical users might think that the system hides additional information about the user's profile that could be used to generate recommendations, so showing only the frequencies of the mentions of the user's aspects may not be enough to satisfy their curiosity and need for further information.

Overall, while most users reported they found the information about others' opinions in explanations useful, the opposite was the case for the information about own preferences, with only a minority of users reporting they found this section useful, and reporting comments in this sense, e. g. "It makes sense that a program would analyse my past comments to find out about my preference...", or "It could be more useful if there was an explanation of how my preferences are used in the calculation" (the latter by a user assigned to a non user preferences condition in Experiment 1). While the difficulty in understanding this information seems to play an important role in this regard, as discussed above, we believe that in the face of a lack of motivation or "feeling of personal relevance" to perform the task, and the need for greater cognitive effort to do so, the user may simply choose not to attend this section, as discussed by [7, 41].

Overall, the results suggest that users seemed to be much more interested in other people's opinion and their weight in the recommendation, rather than how these recommendations fit their own preferences. The reasons for this could be twofold: 1) domain under study is an experience good, where the search for information is characterized by a greater reliance on word-of-mouth [29, 24], and where users might be interested, for example, in finding aspects that had a prominent negative opinion, even when the aspect is not necessarily the most important for them. 2) user models enabled by methods like EFM might not accurately reflect users' real preferences.

As for the explanatory model chosen as inspiration for our study, we believe that the user profile obtained using methods such as the Explicit Factors Model (EFM) [53], may not fully reflect the true preferences of the user, as addressing an aspect in a review, other than reflecting one's own preference, may be motivated by other factors. On the one hand, customers report on the aspects they consider satisfactory or unsatisfactory, but the nature of these aspects may define the satisfaction report on them, as discussed by Chowdhary and Prakas [11]: the presence of some aspects that are taken for granted (cleanliness, for example) may not lead to customers' satisfaction, while their absence leads to dissatisfaction and subsequent reporting. Similarly, motivational factors (e. g., proximity to the beach) can lead customers to satisfaction, but their absence does not necessarily cause a negative report.

On the other hand, when inspecting the data we aimed to provide to our participants in the experimental set-up, we observed that in many cases, users in dataset had fairly homogeneous frequencies of reporting aspects in their reviews i. e., many of them tend to talk about general aspects (e. g., "room", "facilities") in similar proportions. This makes it difficult, in some cases, to detect compelling preferences, which can be prominently represented in an explanation. Thus, we believe that if all aspects have a very similar assessment of relevance (and thus the bars or numbers in the chart look almost the same) the preference information in explanations might be perceived as irrelevant, unnecessary, and even confusing to users. This seemed to be the case for one of the study participants, assigned to the condition bar chart – user preferences displayed, who reported: "I did not understand the left side of the graph which was consistent across about the features relevant to me (seems weird and confusing to include that)". In this regard, however, further evidence is needed to confirm that this is actually the case.

6.2 Effect of Presentation Style

In regard to presentation style, we compared users' perception of explanations consisting of tables or bar charts, that provided an aggregated view of positive and negative opinions given by users to every hotel. Here, we did not find a salient preference of one style over the other. Additionally, despite no significant interaction effect between visualization familiarity and display style was found, we observed, in line with our **H2**, that visualization familiarity might play a role in this perception, since users with higher scores in relation to this user characteristic, gave higher usefulness scores to the bar charts as part of the explanations.

Additionally, our results suggest an interaction effect between rational decision-making style and presentation style on the perception of explanations,so that users with a more rational decision-making style reported higher confidence scores for explanations consisting of tables, while the opposite trend was observed for bar chart explanations. This could be explained by the tendency to seek more detailed information when making decisions, which characterizes individuals with a predominantly rational decision-making style, who may be more interested in evaluating explicit and accurate numbers (such as those presented in the table), compared to less rational users, who may benefit more from representations that allow faster comparisons (such as the bar chart). In this respect, according to Spence and Lewandowsky [36], a presentation of data by means of a table may be more beneficial than the use of a graphical representation, when the objective is the evaluation of exact numbers, and provided that the number of data points presented remains low (in our case it is 5, the number of aspects for which information is provided). The above is also consistent with another of our findings, in which users with a predominantly intuitive decision-making style reported significantly higher scores on the usefulness of bar charts, which seems to be a consequence of the rapid processing of information enabled by graphical representations.

Overall, and despite the differences in perception between tables and bar charts in terms of user characteristics, most users found the two types of explanatory components to be useful, and although the perception of usefulness of the bar chart is slightly more positive than that of the table, this difference is not significant, so we can conclude that both types of presentation are useful to users.

6.3 Main Effect of User Characteristics

So far we have discussed how user characteristics mediate the effect of user preferences or display style on the perception of both the system and its explanations. However, it is important to note that we also observed main effects of decision-making style and social awareness on participants' perception. In particular, we observed that users with a predominant rational style seemed to perceive a greater benefit of the explanation in helping them make faster and better decisions, and as a good means to believe that the recommender is honest, while more intuitive users reported a more positive perception on the quality of explanations, i. e. they like it better and found them more relevant. We believe that the reason why more rational users did not necessarily like our explanations much more could be the lack of additional and detailed arguments addressing the causes of the positive and negative evaluation by customers, given their tendency to examine the information in depth when making decisions, while more intuitive users do not need to go into such detail, and can be satisfied with the aggregate view of opinions we provide in our proposed design. In fact, we received several observations in this regard: "Written reviews from others could be helpful. Rather than just the amount of positive or negative opinions, if you could see specificaly (sic) why they rated the hotel that way it would help personalize your experience even more.", "I think specific comments and reviews would've been helpful in making a final decision. I prefer to read other users' comments about their hotel stay to make a more informed decision".

Furthermore, our results also suggest that social awareness seems to play a significant role in the perception of review-based RS, since we found significant main effects of social awareness on almost all variables evaluated, which seems to be a natural consequence of using users' opinions as a basis for generating explanations, which seems to benefit greatly people with a more pronounced tendency to listen to others and take their opinions into account.

6.4 Usefulness of Origin of Information

Finally, with respect to our **H3**, we found that participants did not find indications of the origin of the information significantly more useful, unless user characteristics such as social awareness were taken into account. In this case, users who were more willing to consider other opinions found more useful the explanatory component reporting the explanations' source of information (i. e. the reviews written by users). In this regard, it is possible that users with less social awareness, being less interested in others' opinions, might have been disappointed because of the expectation that other sources of information would be taken into account when generating recommendations. We believe that this mismatch between the user's conceptual model and the transmitted system's conceptual [21] model could have resulted in a lower usefulness score for this section. However, an alternative explanation could be that users found that information redundant, which could be the case for users who felt that the information on the origin of the information represented in the explanation was already sufficiently self-explanatory.

6.5 Limitations

An important limitation of our study is the fact that user's preferred aspects were fixed and participants were instructed to pretend that those aspects were the ones that mattered most to them, aiming to give a practical work around to the cold-start problem in the user study design. However, we acknowledge that this might interfere with the real perceived benefit of providing the user preferences as part of the explanations.

Additionally, we acknowledge that the use of the Amazon Mechanical Turk implies an important challenge in regard to high quality responses. Here, despite our implemented quality control and the bonus offered, further actions might be still evaluated, aiming to encourage users to genuinely make a decision. In this case, a game strategy could be used, in which users are asked to solve a specific task, for example, to choose the hotel that fits certain conditions using the information provided in the explanations, and to receive a bonus only if the task is solved successfully.

7 Conclusion and Future Work

In this paper, we have presented the design of argumentative explanations based on reviews, in display styles that involve visual representations like tabulated data and bar charts, as well as information about the user preferences. We also addressed the role that individual differences regarding decision making styles, social awareness and visual familiarity play in such perception. Although we found no main differences in perception between the regarded display styles, nor the presence or absence of user preferences in explanations, we found that, when taking into account user characteristics, i. e. social awareness, rational or intuitive decision making style, we are able to do detect differences in explanations' perception between users.

Given the variability of perception of explanatory components when taking into account user characteristics, and given the difficulties (even impossibility) posed by a request or automatic inference of them, we suggest explanation designers to consider a more flexible approach, that allow users to interactively request for different presentation styles and explanatory components whenever it is needed. For example, the system could offer an initial view of explanatory information using a chart, and provide an option to visualize the same data as explicit numbers in a table, or within verbal sentences, to ensure that users who require more support to interpret the explanations have the opportunity to do so.

As part of our future work, and in order to mitigate our limitation regarding the use of real user preferences, we plan to provide a mechanism that allows participants to read explanations that fit better to their real preferences, e. g. to request participants preferences and calculate similarity with users within the dataset, so that we obtain the most similar user in terms of preferences, and use them as a proxy to calculate rating predictions.

Furthermore, we plan to work on and test improvements of the explanatory component of user profile, in order to rule out the difficulty of understanding this type of information as the main cause of its lack of usefulness, so that we can further explore the convenience of using reviews as the primary source for modeling user preferences in review-based explanatory methods.

Funding: German Research Foundation (DFG), Grant Number: GRK 2167, Research Training Group "User-Centred Social Media".

References

[1] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 717–725.

[2] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend?: User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 287–300.

[3] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of the Workshop on the Next Stage of Recommender Systems Research, Beyond Personalization IUI 05*. 13–18.

[4] J. Anthony Blair. 2012. The Possibility and Actuality of Visual Arguments. in: Tindale C. (eds), *Groundwork in the Theory of Argumentation 21*, 205–223.

[5] Jacob A. Burack, Tara Flanagan, Terry Peled, Hazel M. Sutton, Catherine Zygmuntowicz, and Jody T. Manly. 2006. Social Perspective-Taking Skills in Maltreated Children and Adolescents. *Developmental Psychology* 42, 2, 207–217.

[6] Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2013. Multi document summarization of evaluative text. In *Computational Intelligence*, Vol. 29. 545–574.

[7] Richard L. Celsi and Jerry C. Olson. 1988. The Role of Involvement in Attention and Comprehension Processes. *Journal of Consumer Research* 15, 2, 210–224.

[8] Michael Chandler. 1973. Egocentrism and Antisocial Behavior: The Assessment and Training of Social Perspective-Taking Skills. *Developmental Psychology* 9, 3, 326–332.

[9] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee*. 1583–1592.

[10] Li Chen and Feng Wang. 2017. Explaining Recommendations Based on Feature Sentiments in Product Reviews. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces – IUI 17*. 17–28.

[11] Nimit Chowdhary and Monika Prakas. 2005. Service Quality: Revisiting the two factors theory. *Journal of Services Research* 5, 1, 61–75.

[12] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic Generation of Natural Language Explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 57:1–57:2.

[13] Paul T. Costa and Robert R. McCrae. 1992. Four ways five factors are basic. *Personality and Individual Differences* 13, 6, 653–665.

[14] Ruihai Dong, Michael P. O Mahony, and Barry Smyth. 2014. Further Experiments in Opinionated Product Recommendation. In *Case Based Reasoning Research and Development*. Springer International Publishing, 110–124.

[15] Michael J. Driver, Kenneth E. Brousseau, and Phil L. Hunsaker. 1990. The dynamic decision maker.

[16] Andrew S. C. Ehrenberg. 1975. *Data reduction: Analyzing and interpreting statistical data*. Wiley, New York.

[17] Collaborative for Academic Social and Emotional Learning. 2013. 2013 CASEL guide: Effective social and emotional learning programs – Preschool and elementary school edition.

[18] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4, 367–382.

[19] Justin Scott Giboney, Susan A Brown, Paul Benjamin Lowry, and Jay F Nunamaker Jr. 2015. User Acceptance of Knowledge-Based System Recommendations: Explanations, Arguments, and Fit. *Decis. Support Syst* 72, 1–10.

[20] Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The Development and Validation of the Rational and Intuitive Decision Styles Scale. *Journal of Personality Assessment* 98, 5, 523–535.

[21] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[22] Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*.

[23] John R Kirby, Phillip J Moore, and Neville J Schofield. 1988. Verbal and visual learning styles. *Contemporary Educational Psychology* 12, 2, 169–184.

[24] Lisa Klein. 1998. Evaluating the Potential of InteractiveMedia through a New Lens: Search versus Experience Goods. In *Journal of Business Research*, Vol. 41. 195–203.

[25] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. In *User Modeling and User-Adapted Interaction*. 441–504.

[26] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI 19)*. ACM, 379–390.

[27] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. In *Information Systems Research*, Vol. 13.

[28] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Intelligent User Interfaces (IUI 16)*, Vol. 2. 256–260.

[29] Philip J. Nelson. 1981. Consumer Information and Advertising. In *Economics of Information*. 42–77.

[30] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model User-Adap* 27, 393–444.

[31] Rosemary Pacini and Seymour Epstein. 1999. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology* 76, 972–987.

[32] Richard E. Petty and John T. Cacioppo. 1986. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer-Verlag, New York.

[33] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems – RecSys 11*. 157–164.

[34] Wolfgang Schnotz. 2014. Integrated Model of Text and Picture Comprehension. In *The Cambridge Handbook of Multimedia Learning (2nd ed.)*. 72–103.

[35] Janet A. Sniezek and Timothy Buckley. 1995. Cueing and Cognitive Conflict in Judge Advisor Decision Making. *Organizational Behavior and Human Decision Processes* 62, 2, 159–174.

[36] Ian Spence and Stephan Lewandowsky. 1991. Displaying proportions and percentages. 5, 1, 61–77.

[37] Paul Thagard and Abninder Litt. 2000. Models of Scientific Explanation. *The Cambridge handbook of computational cognitive modeling*.

[38] Nava Tintarev. 2007. Explanations of recommendations. *Proceedings of the 2007 ACM conference on Recommender systems, RecSys 07*, 203–206.

[39] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 399–439.

[40] Marko Tkalcic and Li Chen. 2015. Personality and Recommender Systems. In *Recommender Systems Handbook*, 715–739.

[41] Peter Todd and Izak Benbasat. 1999. Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection. *Information Systems Research* 10, 4, 356–374.

[42] Stephen E. Toulmin. 1958. The Uses of Argument.

[43] Edward R. Tufte. 1983. *The visual display of quantitative information*. Graphics Press, Cheshire.

[44] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent User Interfaces*. ACM, 47–56.

[45] Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *15th International Conference on Intelligent Text Processing and Computational Linguistics*. 115–127.

[46] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 18*. 165–174.

[47] Rainer Wasinger, James Wallbank, Luiz Pizzato, Judy Kay, Bob Kummerfeld, Matthias Böhmer, and Antonio Krüger. 2013. Scrutable User Models and Personalised Item Recommendation in Mobile Lifestyle Applications. In *User Modeling, Adaptation, and Personalization, UMAP*. 77–88.

[48] Duane T. Wegener, Richard E. Petty, Kevin L. Blankenship, and Brian Detweiler-Bedell. 2010. Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Consumer Psychology* 20, 5–16.

[49] Yao Wu and Martin Ester. 2015. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Eighth ACM International Conference on Web Search and Data Mining*. ACM, 153–162.

[50] Bo Xiao and Izak Benbasat. 2007. ECommerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly* 31, 1, 137–209.

[51] Ilan Yaniv and Maxim Milyavsky. 2007. Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* 103, 104–120.

[52] Markus Zanker and Martin Schoberegger. 2014. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*. 33–36.

[53] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*. 83–92.

Bionotes

Diana C. Hernandez-Bocanegra
University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany
diana.hernandez-bocanegra@uni-due.de

Diana C. Hernandez-Bocanegra is a research associate and PhD student in the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen, and a member of the Interactive Systems Research Group. Her research interests are focused on recommender systems, explainable AI and the exploration of deep learning and NLProc techniques to improve human-computer interaction.

Jürgen Ziegler
University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany
juergen.ziegler@@uni-due.de

Jürgen Ziegler is a full professor in the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen where he directs the Interactive Systems Research Group. His main research interests lie in the areas of human-computer interaction, human-AI cooperation, recommender systems, information visualization, and health applications. Among other scientific functions he is currently editor-in-chief of i-com - Journal of Interactive Media and chair of the German special interest group on user-centred artificial intelligence.

Paper 3

The following paper is reused from:

- Hernandez-Bocanegra, D.C., & Ziegler, J. (2021). Effects of Interactivity and Presentation on Review-Based Explanations for Recommendations. In: Ardito C. et al. (eds) Human-Computer Interaction – INTERACT 2021. INTERACT 2021. Lecture Notes in Computer Science, vol 12933. Springer, Cham. https://doi.org/10.1007/978-3-030-85616-8_35

Acknowledgment:

- Material from Hernandez-Bocanegra, D.C., & Ziegler, J., proceedings of the 18th IFIP TC 13 International Conference on Human-Computer Interaction, published in 2021, Springer, Cham.

Effects of Interactivity and Presentation on Review-Based Explanations for Recommendations

Diana C. Hernandez-Bocanegra( ) and Jürgen Ziegler

University of Duisburg-Essen, 47057 Duisburg, Germany
{diana.hernandez-bocanegra,juergen.ziegler}@uni-due.de

Abstract. User reviews have become an important source for recommending and explaining products or services. Particularly, providing explanations based on user reviews may improve users' perception of a recommender system (RS). However, little is known about how review-based explanations can be effectively and efficiently presented to users of RS. We investigate the potential of interactive explanations in review-based RS in the domain of hotels, and propose an explanation scheme inspired by dialogue models and formal argument structures. Additionally, we also address the combined effect of interactivity and different presentation styles (i.e. using only text, a bar chart or a table), as well as the influence that different user characteristics might have on users' perception of the system and its explanations. To such effect, we implemented a review-based RS using a matrix factorization explanatory method, and conducted a user study. Our results show that providing more interactive explanations in review-based RS has a significant positive influence on the perception of explanation quality, effectiveness and trust in the system by users, and that user characteristics such as rational decision-making style and social awareness also have a significant influence on this perception.

Keywords: Recommender systems · Explanations · Interactivity · User study · User characteristics

1 Introduction

Explaining the recommendations generated algorithmically by a recommender system (RS) has been shown to offer significant benefits for users with respect to factors such as transparency, decision support, or trust in the system [55,56]. Many approaches to explaining the products or services suggested by an RS have been based on ratings provided by other users or properties of the recommended items, approaches related to collaborative and content-based filtering methods [25,58]. More recently, fueled by the advances in natural language processing, user-written reviews have received considerable attention as rich sources of information about an item's benefits and disadvantages, which can be utilized for explanatory purposes. Reviews are, however, subjective, and may be inconsistent with the overall rating given by the user. Even when overcoming the

Published by Springer Nature Switzerland AG 2021
C. Ardito et al. (Eds.): INTERACT 2021, LNCS 12933, pp. 597–618, 2021.
https://doi.org/10.1007/978-3-030-85616-8_35

challenge of processing noisy review texts, the question of which review-based information to show and how to present it is still largely open, partly due to the lack of empirical findings on how to best present review-based explanations, just as there is a general lack of user-centric evaluations of explanatory RS [44].

While as yet no overall theoretical model of explainable recommendations has been established, we propose to analyze explanations through the lens of argumentation theory which has produced a wide range of models of argumentation [5]. One class of these models defines logical structures with elements such as claims, or evidence to support or refute claims. A second class of models [63] have abandoned the idea of static argumentation models and propose a dialectical approach, focusing on the exchange of arguments within a dialogue between two parties. This approach led to the formulation of dialogue models of explanation [27,38,62], taking into account the social aspect of the explanatory process (an explainer transfers knowledge to an explainee [40]), which could facilitate the interactive provision of explanatory information in the form of a question-and-answer exchange. However, the practical application of dialogue models in explainable RS and their actual benefit from the users' perspective is yet to be determined.

Thus, grounding on argumentation theory and dialogue models of explanation, we formulated and tested an interactive approach to explanations based on reviews, that facilitates the exploration of arguments that support claims made by the system (i.e. an item is worth purchasing), while providing answers to some of their potential questions at different levels of detail (e.g. what was reported on [feature]?). To this end, we adopted the definition of interactivity by Steuer [54]: "extent to which users can participate in modifying the form and content of mediated environment in real time", and characterized the degree of interactivity of explanations through the Liu and Shrum dimensions of interactivity [35]: *active control* and *two-way communication*. The first is characterized by voluntary actions that can influence the user experience, reflected in our proposal by the possibility to use hyperlinks and buttons, that allow users to navigate explanatory information at will. The second refers to the ability of two parties to communicate to one another, reflected in our proposal by the ability to indicate the system which are their most relevant features, so the answers are adjusted accordingly.

While interactive explanations have been already addressed in the field of explainable artificial intelligence (XAI), their impact in explainable RS remains largely unexplored, as well as the empirical validation of their effects on users. Hence, we aimed to provide empirical evidence of the effect that an implementation of this approach may have on users' perception. More specifically, we evaluated users' perception in terms of the quality of explanations, and of the explanatory objectives: transparency, effectiveness and trust, as defined by [55], and aim to answer: **RQ1**: How do users perceive review-based explanations with different degrees of *interactivity*, in terms of explanation quality, and of the transparency, efficiency and trust in the system? We also aimed to test the combined effect of explanation interactivity and different presentation styles, particularly:

using only text, using a bar chart or using a table, to show, among others, the distribution of positive and negative comments on the quality of an item. Here, we were interested to inquire, for example, whether users who find a presentation style less satisfactory might benefit from interactive options that allow them to clarify their doubts. Thus: **RQ2**: How do different *presentation styles* influence users' perception of review-based explanations with different degrees of interactivity?

Furthermore, we addressed the influence that different user characteristics might have on the perception of the proposed approach. Regardless of its type, an explanation may not satisfy all possible explainees [52]. Moreover, individual user characteristics can lead to different perceptions of a RS [30,67], for which we assumed that this would also be the case for explanations, as discussed by [6,26,31]. Since a main objective of providing explanations is to support users in their decision-making, investigating the effect of different personal styles to perform such a process is of particular interest to us. Particularly, we focus on the moderating effect of the *rational* and *intuitive* decision making styles [24], the former characterized as a propensity to search for information and evaluate alternatives exhaustively, and the latter by a quick processing based mostly on hunches and feelings. Furthermore, since review-based explanations rely on the expressed opinions of other users, we investigated the effects of the extent to which users are inclined to adopt the perspective of others when making decisions, a trait defined as *social awareness* by [10]. We also considered *visualization familiarity*, i.e. the extent to which a user is familiar with graphical or tabular representations of information. Consequently, **RQ3**: How do individual differences in decision-making styles, social awareness or visualization familiarity moderate the perception of review-based explanations with different degrees of interactivity and presentation styles?

To address our research questions, we conducted a user study taking as example the hotels domain, since it represents an interesting mix between search goods (with attributes on which complete information can be found before purchase [42]) and experience goods (which cannot be fully known until purchase [42]). Such a product evaluation could benefit from third-party opinions [29,42], potentially rich in argumentative information that can be used for explanatory purposes.

Finally, the contributions of this paper can be summarized as follows:

- We formulate a scheme for explanations as interactive argumentation in review-based RS, inspired by dialogue models and argument structures.
- To test our research questions, we implemented an interface based on the proposed scheme, and a RS based on a matrix factorization model (i.e. EFM, [70]), and sentiment-based aspect detection, using the state of art natural language processing model BERT ([15]).
- We provide empirical evidence of the effect of review-based interactive explanations on users' perception, as well as the influence of user characteristics on such perception.

2 Related Work

Next, we will review work related to review-based explanations, interactive explanations in both explainable artificial intelligence (XAI) and RS, the use of dialogue models in contrast to static models of explanations, and the moderating effect of user characteristics on the perception of explainable RS.

Review-Based Explanations. Review-based explanatory methods leverage user generated content, rich in detailed evaluations on item features, which cannot be deduced from the general ratings, thus enabling the generation of more detailed explanations, compared to collaborative filtering (e.g. "Your neighbors' ratings for this movie" [25]) and content-based approaches (e.g. [58]). Review-based methods allow to provide: **1)** verbal summaries of reviews, using abstractive summarization from natural language generation (NLG) techniques [8,14], **2)** a selection of helpful reviews (or excerpts) that might be relevant to the user, detected using deep learning techniques and attention mechanisms [11,17], **3)** a statistical view of the pros and cons of item features, usually using topic modelling or aspect-based sentiment analysis [16,66,70], information that is integrated to RS algorithms like matrix or tensor factorization [4,64,70]) to generate both recommendations and aspect-based explanations.

Our evaluation is based on the third approach, and is particularly related to the model proposed by [70], since it facilitates getting statistical information on users' opinions, which has been proven to be useful for users [26,41], and can be provided in explanations with different presentation styles (strictly verbal or visual). Yet, the optimal way of presenting explanatory information, either in a textual (short summaries) or a graphical form (e.g. bar charts) remains unclear.

Interactive Explanations. In addition to display factors, a second factor could also influence users' perception of the explanations: the possibility of interacting with the system, to better understand the rationale for its predictions. Interactive explanations have been already addressed in the field of explainable artificial intelligence (XAI) (although to a much lesser extent compared to static explanations [1]). Here, the dominant trend has been to provide mechanisms to check the influence that specific features, points or data segments may have on final predictions of machine learning (ML) algorithms, as in [13,32,51]. However, the impact of such interactive approaches in explainable RS remains largely unexplored. More specifically, the dominant ML interactive approach differs from ours in at least two ways: 1) we use non-discrete and non-categorical sources of information, subjective in nature and unstructured, which, however, can be used to generate both textual and visual structured arguments 2) such approach is designed to meet the needs of domain experts, i.e. users with prior knowledge of artificial intelligence, while we aim to target the general public.

Effects of interactivity have been studied widely in fields like online shopping and advertising [35,53], and more specifically in the evaluation of critique-based RS, where users are able to specify preferences for the system to recalculate recommendations, which has been found to be beneficial for users [12,36,37]. Despite the intuitive advantages that interactivity can bring, interactivity does not always translate into a more positive attitude towards the system, since it also depends on the context and the task performed [35]. Nevertheless, it has also been shown that higher active control is beneficial in environments involving information needs, and a clear goal in mind [35], which is actually our case (i.e. deciding which hotel to book).

Dialogue Models of Explanation. To formulate and test our proposal, we set our focus on argumentative models that may enable the two-way communication desideratum. In contrast to static approaches to explanation, dialogue models have been formulated conceptually [2,38,47,60], allowing arguments over initial claims in explanations, within the scope of an interactive exchange of statements. Despite the potential benefit of using these models to increase users' understanding of intelligent systems [40,65], their practical implementation in RS (and in XAI in general) still lacks sufficient empirical validation [38,40,52]. This dialogical approach contrasts with other argumentative - though static - explanation approaches [3,9,26,33,69] based on static schemes of argumentation (e.g. [23,57]), where little can be done to indicate to the system that the explanation has not been fully understood or accepted, and that additional information is still required.

User Characteristics. We hypothesized (in line with [35]) that a number of user characteristics may moderate the effect of interactive functionalities, on the perception of explanations. Particularly, we aimed to test the moderating effect of decision-making styles and social awareness. In regard to the former, research has shown that it is determined significantly by preferences and abilities to process available information [19]. Particularly, we believe that users with a predominant rational decision making style would better perceive explanations with a higher degree of interactivity, than explanations with less possibility of interaction, given their tendency to thoroughly explore information when making decisions [24]. On the other hand, more intuitive users may not find the interactive explanations very satisfactory, given their tendency to make decisions through a quicker process [24], so that a first explanatory view would be sufficient, and it would not be necessary to navigate in depth the arguments that the system can offer. Here, [26] noted that rationality and intuition might not be diametrically opposed constructs. In their study, although most participants reported that they thoroughly evaluate the available information when making decisions, many of them also reported a tendency to use their intuition.

As for social awareness, and in line with results reported by [26], we hypothesize that users with a higher social awareness may perceive explanations with higher interactivity more positively, given their tendency to take into account the opinions of others, and to adjust their own using those of others, while choosing between various alternatives [50], which has been proved to be beneficial during decision making [68], and is facilitated by our approach.

Finally, in regard to presentation styles, visual arguments (a combination of visual and verbal information) may have a greater "rhetorical power potential" than verbal arguments, due (among others) to their greater immediacy (possibility of quick processing) [7]. This could especially benefit users with a predominantly intuitive decision-making style, due to their usually quick manner of making decisions, based mostly on first impressions [24]. However, users with lower visual abilities might benefit less from a presentation based on images or graphics [28,49]. Consequently, we believe that when exposed to graphic-based explanation formats, higher interactive explanations may be beneficial to users with lower visual familiarity, as they could access additional information to better understand the explanations provided.

3 Scheme for Explanations as Interactive Argumentation in Review-Based RS

To evaluate our research questions, we designed an interaction scheme for the exploration of explanatory arguments in review-based RS. A recommendation issued by a RS can be considered a specific form of a claim, namely that the user will find the recommended item useful or pleasing [18]. The role of an explanation is thus to provide supportive evidence (or rebuttals) for this claim. Claims are, however, also present in the individual user's rating and opinions, which may require explaining their grounds as well, thus creating a complex multi-level argumentative structure in an explainable RS, a concern also raised in [22]. To formulate an explanation scheme able to support this type of structure, we considered dialog-based explanation models [38,61,62], in which instead of a single issue of explanatory utterances, an explanation process is regarded as an interaction, where a user could indicate when additional arguments are required, to increase their understanding of system claims.

In this context, Walton [61,62] modeled explanation requests (user questions) and explanation attempts (a set of assertions as system response). On the other hand, Madumal et al. [38] noted that argumentation may occur within explanation, and modeled the shift between explanatory and argumentative dialog, as well as the explanatory loops that can be triggered, when follow-up questions arise. While this type of models may help to define the moves allowed within an explanatory interaction, they offer little indication of how the arguments within the interaction moves should be structured, to increase their acceptance by users. To this end, we rely on the scheme by Habernal et al. [23], an adaptation of the Toulmin model of argumentation [57], formulated to better represent the kind of arguments usually found in user-generated content. This scheme involves: claim (conclusion of the argument), premise (a general reason to accept a claim), backing (specific information or additional evidence to support the claim), rebuttal (statement that attacks the claim) and refutation (statement that attacks the rebuttal).

Our proposed scheme is shown in Fig. 1. Unlike Walton, who modeled explanatory movements as explanation requests and attempts, we considered an explanation process as a sequence of *argumentation attempts* (the system intends to provide arguments to explain something) followed by *argument requests* (the user ask the system to provide - follow-up - arguments that support the claim that user will find the recommended item useful). A missing element of Walton's model in our scheme is a feedback mechanism so that users can indicate whether

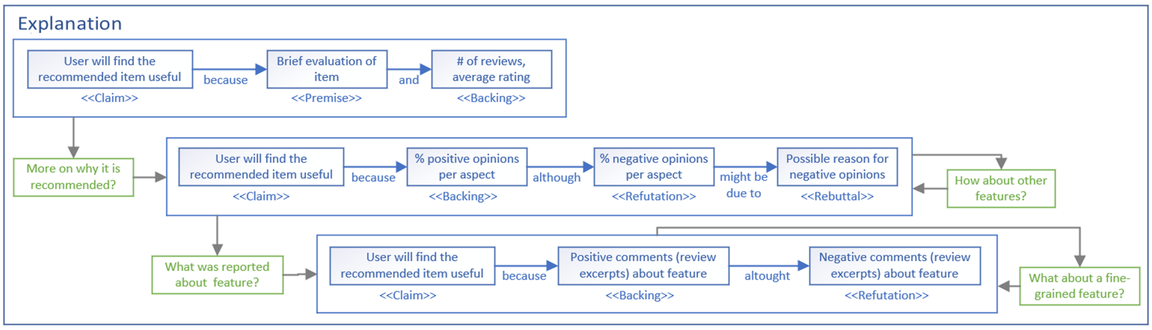
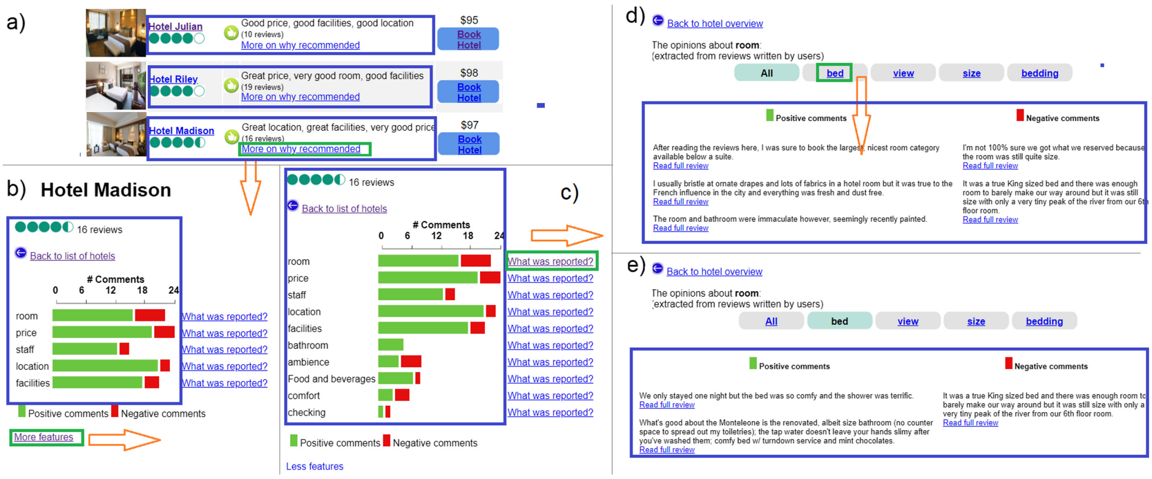

Explanation
User will find the recommended item useful
<<Claim>>
because
Brief evaluation of item
<<Premise>>
and
# of reviews, average rating
<<Backing>>
More on why it is recommended?
User will find the recommended item useful
<<Claim>>
because
% positive opinions per aspect
<<Backing>>
although
% negative opinions per aspect
<<Refutation>>
might be due to
Possible reason for negative opinions
<<Rebuttal>>
How about other features?
What was reported about feature?
User will find the recommended item useful
<<Claim>>
because
Positive comments (review excerpts) about feature
<<Backing>>
altought
Negative comments (review excerpts) about feature
<<Refutation>>
What about a fine-grained feature?

Fig. 1. Scheme for explanations as interactive argumentation in review-based RS. Blue boxes: argumentation attempts by the system, green boxes: argument requests by users. (Color figure online)

full understanding has been achieved, which is left for future work. The realization of our scheme as an user interface is depicted in Fig. 2. Here, design features like links and buttons enable argument requests by users, e.g.: the link "what was reported?" (Fig. 2b) fosters the interactive features *active control* (control on what aspects the system should focus on in its argumentation) and *two-way communication* (user communicates to the system that further argument backing is needed), which triggers a system argumentation attempt.



Good price, good facilities, good location (10 reviews)

Great price, very good room, good facilities (19 reviews)

Great location, great facilities, very good price (16 reviews)

The opinions about **room**:
(extracted from reviews written by users)

After reading the reviews here, I was sure to book the largest, nicest room category available below a suite.

I usually bristle at ornate drapes and lots of fabrics in a hotel room but it was true to the French influence in the city and everything was fresh and dust free.

The room and bathroom were immaculate however, seemingly recently painted.

I'm not 100% sure we got what we reserved because the room was still quite size.

It was a true King sized bed and there was enough room to barely make our way around but it was still size with only a very tiny peak of the river from our 6th floor room.

The opinions about **room**:
(extracted from reviews written by users)

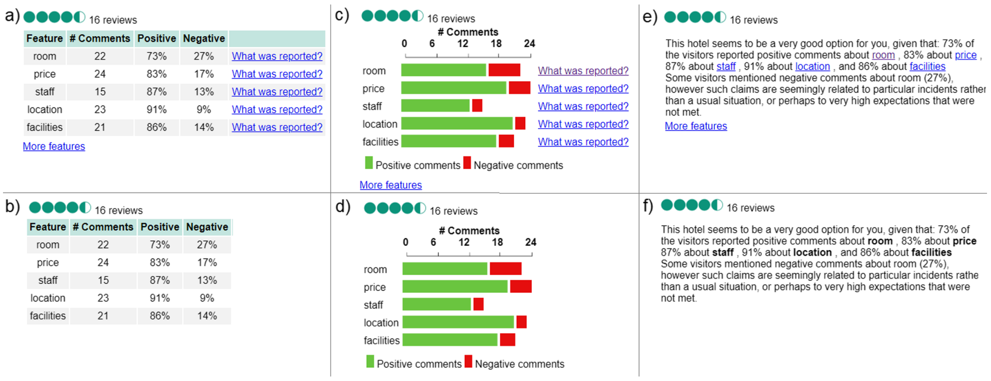
We only stayed one night but the bed was so comfy and the shower was terrific.

What's good about the Monteleone is the renovated, albeit size bathroom (no counter space to spread out my toiletries); the tap water doesn't leave your hands slimy after you've washed them; comfy bed w/ turndown service and mint chocolates.

It was a true King sized bed and there was enough room to barely make our way around but it was still size with only a very tiny peak of the river from our 6th floor room.

Fig. 2. Screenshots of system used in user study. Enclosed in blue: argumentation attempts; in green: argument requests. Orange arrows: sequence of allowed moves, pointing to the next interface. a) List of recommended items; clicking on "More on why recommended" displays: b–c) aggregation of comments by aspect; clicking on "What was reported?" displays: d) comments on chosen aspect; clicking a fine-grained feature button, displays: e) comments on chosen feature. c, d, e enabled only in study condition interactivity "high". (Color figure online)

An explanatory dialogue can take place both through verbal interactions and through a visual interface (non-verbal communication, or a combination of verbal and visual elements) [38,40]. As for presentation, while arguments are usually associated with oral or written speech, arguments can also be communicated using visual representations (e.g. graphics or images) [7]. Thus, we considered the following styles for the argumentation attempt "% of positive and negative opinions": 1) Table (Fig. 3a, 3b), bar chart (Fig. 3c, 3d), and text (Fig. 3e, 3f), the latter using the template proposed by [26], which facilitates the display of rebuttal statements, which can hardly be represented graphically.



a) ●●●●◐ 16 reviews

Feature	# Comments	Positive	Negative	
room	22	73%	27%	What was reported?
price	24	83%	17%	What was reported?
staff	15	87%	13%	What was reported?
location	23	91%	9%	What was reported?
facilities	21	86%	14%	What was reported?

More features

b) ●●●●◐ 16 reviews

Feature	# Comments	Positive	Negative
room	22	73%	27%
price	24	83%	17%
staff	15	87%	13%
location	23	91%	9%
facilities	21	86%	14%

c) ●●●●◐ 16 reviews

d) ●●●●◐ 16 reviews

e) ●●●●◐ 16 reviews

This hotel seems to be a very good option for you, given that: 73% of the visitors reported positive comments about room , 83% about price , 87% about staff , 91% about location , and 86% about facilities Some visitors mentioned negative comments about room (27%), however such claims are seemingly related to particular incidents rather than a usual situation, or perhaps to very high expectations that were not met.

More features

f) ●●●●◐ 16 reviews

This hotel seems to be a very good option for you, given that: 73% of the visitors reported positive comments about **room** , 83% about **price** 87% about **staff** , 91% about **location** , and 86% about **facilities** Some visitors mentioned negative comments about room (27%), however such claims are seemingly related to particular incidents rathe than a usual situation, or perhaps to very high expectations that were not met.

Fig. 3. Manipulation of *presentation* style in combination with *interactivity*, in user study. Left: *table*, middle *bar chart*, right *text*. Top: interactivity *high*, bottom: interactivity *low*.

4 User Study

To answer our research questions, we implemented a RS that reflects the scheme described in Sect. 3, and conducted a user study to compare users' perception of the overall system (in terms of the dependent variables (DVs): transparency, effectiveness and trust), and of the specific aspects of explanations (in terms of the DVs: explanation confidence, transparency, persuasiveness, satisfaction and sufficiency). As independent variables (IVs) we considered the factors *interactivity* and *presentation*. Possible values of IV interactivity are: "high" (users could make all possible argument requests, Fig. 1 and 2), and "low" (users could only make the initial argument request "more on why recommended?"). Possible values of IV presentation are: *table* (Fig. 3 a, b), *bar chart* (Fig. 3 c,d) and *text* (Fig. 3 e,f). The study follows a 3 × 2 between-subjects design, and each participant was assigned randomly to one of six conditions (combination of *interactivity* and *presentation* style). As covariates, we considered the user characteristics: rational and intuitive decision-making style, social awareness and visualization familiarity. We hypothesized:

H1: Users' perception of the system and its explanations is more positive when they are given explanations with higher interactivity.

H2: Users with a predominantly rational decision style perceive explanations with higher interactivity more positively than less rational decision makers.

H3: Less intuitive users perceive explanations with higher interactivity more positively, compared to more intuitive users.

H4: Users with greater social awareness perceive higher interactive explanations more positively than users with less social awareness.

H5a: Users with a predominantly intuitive decision-making style or **H5b** a greater visualization familiarity will prefer bar chart explanations over text explanations, regardless of interactivity.

H6: Users who are less familiar with data visualization will perceive explanations with higher interactivity more positively, particularly in the case of more challenging visualizations such as bar charts.

4.1 Questionnaires

Evaluation: We utilized items from [46] to evaluate the perception of system transparency (construct *transparency*, user understands why items were recommended), of system effectiveness [30] (internal reliability Cronbach's $\alpha = 0.85$, construct *perceived system effectiveness*, system is useful and helps the user to make better choices), and of trust in the system [39] ($\alpha = 0.90$, constructs *trusting beliefs*, user considers the system to be honest and trusts its recommendations; and *trusting intentions*, user willing to share information). We used the user experience items (UXP) of [31] to address explanations reception, which we will refer to as *explanation quality* ($\alpha = 0.82$), comprising: explanation confidence (user is confident that she/he would like the recommended item), explanation transparency (explanation makes the recommendation process clear), explanation satisfaction (user would enjoy a system if recommendations are presented this way), and explanation persuasiveness (explanations are convincing). We added an item adapted from [17] (explanations provided are sufficient to make a decision) to evaluate explanation sufficiency. All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

User Characteristics: We used all the items of the Rational and Intuitive Decision Styles Scale [24] (internal reliability Cronbach's $\alpha = 0.84$ and $\alpha = 0.92$, respectively), the scale of the social awareness competency proposed by [10] ($\alpha = 0.70$), and the visualization familiarity items proposed by [31] ($\alpha = 0.86$). All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

4.2 Participants

We recruited 170 participants (66 female, mean age 37.61 and range between 18 and 72) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 98%, and a number of HITs approved greater than 500. We applied a quality check to select participants with quality survey responses (we asked validation questions to check attentiveness within questionnaires, and questions related to the content of the system, e.g. "recommendations were based on: Opinions of celebrities, True/False", "The purpose of this question is to check attentiveness, please mark Disagree"). We discard participants with less than 10 (out of 12) correct answers, or no effective interaction with the system (checked in logs). The responses of 27 of the 197 initial participants were then discarded for a final sample of 170 subjects, statistical power of 90%, α =0.05, power value above conventional for adequacy of .80 are considered acceptable [45]; An 'a priori' type of analysis was performed in G*power software [21]. Participants were rewarded with $1.4 plus a bonus up to $0.40, depending on the length and number of arguments provided in their response to the question "Why did you choose this hotel?", and the extent to which those arguments referred to explanations and information provided by the system. Time devoted to the task by participants (in minutes): M = 10.88, SD = 1.62.

4.3 Procedure

Instructions indicated that a list of hotels reflecting the results of a hypothetical hotels' search and within the same price range would be presented (i.e. no filters to search hotels were offered to participants), and that they could click on the name of a desired hotel to see general information about it. However, we asked, as we were more interested in their views on the explanations given for each recommendation, to click on the "More on why recommended" links of hotels they might be interested in, and to explore the information provided. No further instructions were given regarding how to interact with the different interaction options, since we were interested to address to what extent the users used them or not. Users were instructed to indicate which hotel they would finally choose, and to write a few sentences reporting their reasons for it, for which a bonus up to $0.4 would be paid, depending on the quality of this response, with the aim of achieving a more motivated choice by the participants, as well as to encourage a more effective interaction with the system. We then presented a cover story, which sought to establish a common starting point in terms of travel motivation (a holiday trip). Next, we presented to the participants the system showing a list of 30 recommended hotels (sorted by predicted rating), and their corresponding personalized explanations (system implementation details in Sect. 4.4). Finally, evaluation and validation questions were presented, plus an open-ended one, asking for general opinions and suggestions about the explanations.

4.4 Dataset and Implemented System

Dataset and Aspect Annotation: ArguAna [59], includes hotel reviews and ratings from TripAdvisor; sentiment and explicit features are annotated sentence wise. We categorized the explicit features in 10 general features (room, price, staff, location, facilities, bathroom, ambience, food and beverages, comfort and checking), with the help of 2 annotators (Krippendorff's alpha of 0.72), aiming to train a classifier to detect the main aspect addressed in a sentence (e.g. "I loved the bedding" would be classified as *room*).

Aspect-Based Sentiment Detection: We trained a BERT classifier [15] to detect the general feature addressed within a sentence: we used a 12-layer model (*BertForSequenceClassification*), 6274 training sentences, 1569 test sentences, F-score 0.84 (macro avg.). We also trained a BERT classifier to detect the sentiment polarity, using a 12-layer model (*BertForSequenceClassification*), 22674 training sentences, 5632 test sentences, F-score 0.94 (macro avg.). Classifier was used to **1)** consolidate the quality of hotels and relevance of aspects to users (see Figs. 2b, 2d), and **2)** to present participants with negative and positive excerpts from reviews regarding a chosen feature (Fig. 2d, 2e).

Explainable RS Method: We implemented the Explicit Factor Model (EFM) [70], a review-based matrix factorization (MF) method to generate both recommendations and explanations. The rating matrix (ratings granted by users to items) consisted of 1284 items and 884 users extracted from the ArguAna dataset (only users with at least 5 written reviews were included), for a total of 5210 ratings. Item quality and user preferences matrices were consolidated using the sentiment detection described previously. Each element of the former matrix measures the quality of the item for each aspect, while the elements of the latter measure the extent to which the user cares about an aspect. The number of explicit features was set to 10. Model-specific hyperparameters were selected via grid-search-like optimization. After 100 iterations, we reached an RMSE of 1.27, a metric used to measure the differences between dataset values and the values predicted by the RS model. Values of predicted rating matrix were used to sort recommendations, and shown within explanations (average hotel rating with 1–5 green circles). Values of quality matrix were used to calculate the percentages of positive and negative comments on aspects (Fig. 3).

Personalization Mechanism: To reduce implications of the *cold start* problem [48] (system does not have enough information about the user to generate an adequate profile and thus, personalized recommendations), participants were asked for the five hotel features that mattered most to them, in order of importance. The system calculated a similarity measure, to detect users within the EFM preference matrix with a similar order of preferences. Then the most similar user was used as a proxy to generate recommendations, i.e. we selected the predicted ratings of this proxy user, and used them to sort recommendations and features within explanations.

4.5 Data Analysis

We evaluated the effect that IVs (interactivity and presentation style) may have on 2 different levels: **1)** *overall system* perception (DVs explanation quality, and system transparency, effectiveness and trust), and **2)** perception of specific aspects of *explanations* (DVs explanation confidence, transparency, satisfaction, persuasiveness and sufficiency), and to what extent the covariates (user characteristics: rational and the intuitive decision making styles, social awareness and visualization familiarity) could influence such perception.

Evaluation scores (DVs' scores) for each individual were calculated as the average of the reported values for the scale items, in case of multi-item scales. Scores on "explanation quality" were calculated for each individual as the average of scores on specific aspects of explanations, and the covariates scores as the average of the reported values for items of every scale. Internal consistency

(Cronbach's alpha) was checked for system evaluation and user characteristics constructs (reported in Sect. 4.1).

Overall System Perception: Given that DVs are continuous and correlated (see Table 1), a MANCOVA analysis was performed. Subsequent ANCOVA were performed to test main effects of IVs and covariates, as well as the effect of interactions between them. Q-Q plots of residuals were checked to validate the adequacy of the analysis.

Perception of Explanations: DVs are ordinal (scores are the reported answers to single questionnaire items), thus we performed ordinal logistic regressions to test influence on DVs by predictor variables (IVs and covariates), no multicollinearity was tested, as well as Q-Q plots of residuals. DVs are also correlated (see Table 2), so significant tests were conducted using Bonferroni adjusted alpha levels of .01 (.05/5).

Use of Interactive Options: Calculated based on system activity logs. A Mann-Whitney U test was used to compare distributions of users characteristics who used or not use such options.

5 Results

5.1 Evaluation and User Characteristics Scores

The average evaluation scores by presentation style and interactivity are shown in Tables 1 and 2. Distributions of the scores of rational (M = 4.35, SD= 0.50) and intuitive (M = 2.59, SD= 0.98) decision making styles, social awareness (M = 4.04, SD= 0.53) and visualization familiarity (M = 3.23, SD= 0.95) are depicted in Fig. 4a.

Table 1. Mean values and standard deviations of perception on the overall system, per *presentation* style and *interactivity* (n = 170), $p < 0.05$*, $p < 0.01$**; values reported with a 5-Likert scale; higher mean values correspond to a positive perception of the overall RS. Pearson correlation matrix, $p < 0.001$ for all coefficients.

Variable	*Presentation*						*Interactivity*				*Correlation*				
	Text		Table		Bar chart		Low		High		Variable				
	M	SD	M	SD	M	SD	M	SD	M	SD	1	2	3	4	
1. Expl. Quality	3.98	0.52	4.10	0.56	4.07	0.70	3.92	0.61	4.17**	0.55					
2. Transparency	4.14	0.52	4.11	0.86	3.91	0.99	4.02	0.78	4.08	0.86	0.51	—			
3. Effectiveness	3.95	0.69	4.05	0.73	4.04	0.78	3.91	0.79	4.11*	0.67	0.67	0.75	0.56	—	
4. Trust	3.91	0.60	3.99	0.58	3.97	0.72	3.86	0.67	4.05*	0.57	0.57	0.74	0.55	0.79	—

5.2 Overall System Perception

Interactivity: We found a significant multivariate effect of interactivity on overall system perception $F(4{,}157) = 2.68$, $p = .034$. Univariate tests revealed that

Table 2. Mean values and standard deviations of perception on explanation specific aspects, per *presentation* style and *interactivity* (n = 170), p < 0.05*, p < 0.01**; values reported with a 5-Likert scale; higher mean values correspond to a positive perception on the explanations. Pearson correlation matrix, p < 0.001 for all coefficients.

Variable	*Presentation*						*Interactivity*				*Correlation*				
	Text		Table		Bar chart		Low		High		Variable				
	M	SD	M	SD	M	SD	M	SD	M	SD	1	2	3	4	5
1. Expl. confidence	4.09	0.55	4.11	0.65	4.05	0.85	3.95	0.74	4.21*	0.62					
2. Expl. transparency	4.16	0.73	4.19	0.83	4.16	0.86	4.05	0.84	4.29*	0.77	0.60	—			
3. Expl. satisfaction	3.84	0.85	4.09	0.79	4.11	0.80	3.88	0.84	4.14*	0.77	0.40	0.53	—		
4. Expl. persuasiveness	3.84	0.71	3.96	0.71	3.93	0.82	3.82	0.71	4.00	0.77	0.64	0.47	0.45	—	
5. Expl. sufficiency	3.96	0.79	4.14	0.81	4.09	0.83	3.89	0.87	4.23**	0.71	0.40	0.44	0.50	0.45	—

interactivity significantly influences the perception of explanation quality $F(1{,}168) = 9.76$, $p = .002$, effectiveness $F(1{,}168) = 4.02$, $p = .047$, and trust $F(1{,}168) = 4.63$, $p = 0.033$. In all these cases, the average of every variable was higher for the *high* condition than for *low* condition (see Table 1).

Presentation Style: We found no significant main effect of *presentation* style.

Rational Decision-Making Style: We found a significant multivariate effect of rational style, $F(4{,}157) = 7.55$, $p < .001$. Univariate tests revealed a main effect of rational decision-making style on explanation quality, $F(1{,}168) = 20.27$, $p < .001$, system transparency $F(1{,}168) = 8.25$, $p = .005$, effectiveness, $F(1{,}168) = 26.76$, $p < .001$ and trust, $F(1{,}168) = 24.94$, $p < .001$. In all these cases, a positive trend was observed between these variables and the rational decision-making style, i.e. the higher the rational decision-making score, the higher the perceived explanation quality, the transparency, the effectiveness and the trust, independent of style or interactivity (Fig. 4b).

Social Awareness: We found a significant multivariate effect of social awareness, $F(4{,}157) = 6.41$, $p < .001$. Univariate tests revealed a main effect of social awareness on explanation quality $F(1{,}168) = 17.25$, $p < .001$, system transparency $F(1{,}168) = 12.57$, $p < .001$, effectiveness $F(1{,}168) = 22.85$, $p < .001$ and trust $F(1{,}168) = 18.02$, $p < .001$. In all these cases, a positive trend was observed between these variables and social awareness, i.e. the higher the social awareness score, the higher the perceived explanation quality, the transparency, the effectiveness and the trust, independent of style or interactivity (Fig. 4c).

5.3 Perception of Explanations

Interactivity: We found a main significant effect of interactivity; here, the odds of participants reporting higher values of explanation sufficiency when interactivity *high* was 2.30 (95% CI, 1.26 to 4.29) times that of interactivity *low*, a statistically significant effect, Wald $\chi2(1) = 7.32$, $p = .007$. We observed a similar pattern in relation to explanation confidence ($p = .017$), explanation transparency ($p = .043$) and explanation satisfaction ($p = .041$). However, this association (despite $p < .05$) is non-significant after Bonferroni correction (corrected $p < 0.01$).

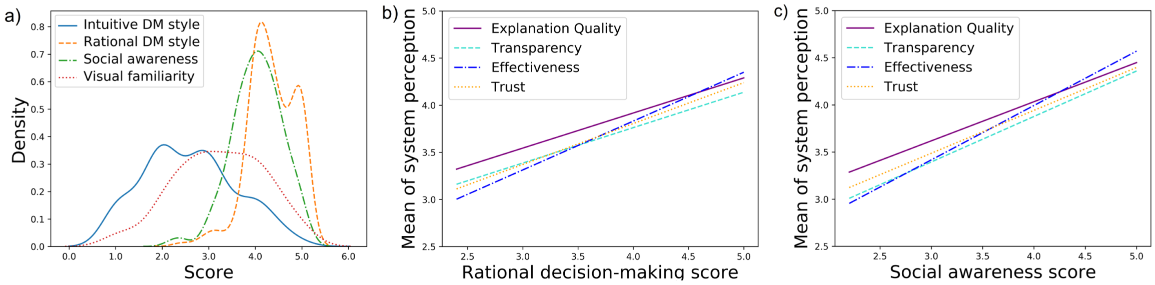


Fig. 4. a) Kernel density estimate of user characteristics scores: rational and intuitive decision making styles, social awareness and visualization familiarity. b) Effect of rational decision-making style on the perception of the overall system (fitted means of individual scores). c) Effect of social awareness on the perception of the overall system (fitted means of individual scores).

Presentation Style: We found no significant main effect of *presentation* style.

Additionally, we observed a possible interaction (p<= 0.05, although non-significant after Bonferroni correction, corrected p< 0.01) between:

Rational Decision-Making Style and Interactivity: An increase in rational decision-making score was associated with an increase in the odds of participants under interactive *high* condition reporting higher values of explanation sufficiency, with an odds ratio of 3.20 (95% CI, 0.99 to 10.65), Wald $\chi 2(1) = 3.81$, p = .051 (Fig. 5a).

Intuitive Decision-Making Style and Presentation Style: An increase in intuitive decision-making score was associated with an increase in the odds of participants under *bar chart* condition reporting higher values of explanation satisfaction, with an odds ratio of 2.40 (95% CI, 1.14 to 5.18), Wald $\chi 2(2) = 5.67$, p = .023, compared to participants under *text* condition (see Fig. 5b).

Social Awareness and Interactivity: An increase in social awareness score was associated with an increase in the odds of participants under interactive *high* condition reporting higher values of explanation persuasiveness, with an odds ratio of 3.83 (95% CI, 1.20 to 12.34), Wald $\chi 2(1) = 5.17$, p = .023 (Fig. 5c).

Visualization Familiarity and Interactivity: An increase in visualization familiarity score was associated with an increase in the odds of participants under interactive *high* condition reporting higher values of explanation satisfaction, odds ratio of 1.91 (95% CI, 1.03 to 3.58), Wald $\chi 2(1) = 4.24$, p = .039 (Fig. 5d).

5.4 Use of Interaction Options

48% of the users assigned to the interactivity *high* conditions used at least one of the interaction options provided. 48.15% of participants used the 'more features' option when explanations were displayed using table, 26.92% using bar chart and 33.3% using text. 55.56% of participants used the 'what was reported' option when explanations were displayed as table, 50% as bar chart and 3.7% as text. And 22.22% of participants used the 'comments on specific features' option when

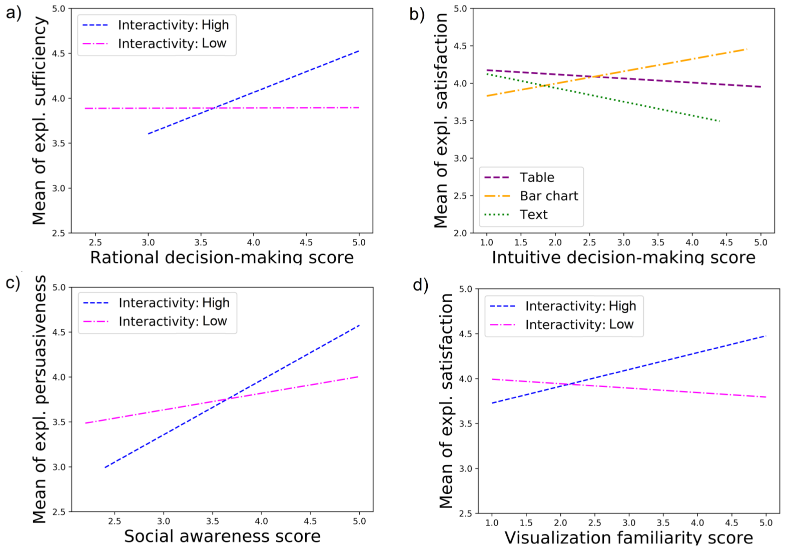


Fig. 5. Interaction plots (fitted means of individual scores) for perception of explanation: a) sufficiency, interaction between interactivity and rational decision-making style. b) satisfaction, interaction between presentation and intuitive decision-making. c) persuasiveness, interaction between interactivity and social awareness. d) satisfaction, interaction between interactivity and visualization familiarity.

explanations were displayed as table, 19.23% as bar chart and 3.7% as text. Additionally, a Mann-Whitney U test revealed that the average of visualization familiarity scores of users who used the interaction options ($M = 2.98$, $SD = 1.05$) is significantly lower than the score of those that did not use them ($M = 3.41$, $SD = 0.85$), $U(\mathrm{N}_{used} = 41, \mathrm{N}_{notused} = 44) = 678.50$, $p = .024$).

6 Discussion

In regard to our **RQ1**, our results show that greater interactivity has a significantly positive effect on users' perception, in terms of system effectiveness and trust, as well as of explanation quality, compared to explanations with lower interactivity, thus confirming our **H1**. We believe that the interactivity aspects addressed in our proposal could play a determining role in the observed effect, namely: active control and two-way communication, fostered in turn by design features such as links and buttons representing argument requests. Active control by enabling users to be in control of which argumentative content to display; two-way communication by enabling them to indicate the system which argumentative statements require further elaboration, and which features are of real relevance at the time of making the decision, an approach that might contribute to a better acceptance and understanding of explanations, as predicted by dialogue models of explanation [27,62].

However, the benefit and actual use of interactive options in review-based explanations might be influenced by individual differences, as discussed by [35] for the scope of online advertising and shopping. In particular, and regarding our

RQ3, we found that the way people process information when making decisions would play an important role in the perception of interactive review-based explanations. More precisely, and in line with **H2**, we found that greater interactivity might have a more positive effect on the perception of explanation sufficiency by more rational users, which is explained by the propensity of people with a predominant rational decision-making style, to search for information and evaluate alternatives exhaustively [24]. However, and contrary to our expectations, we observed that the degree of intuitive decision style did not moderate the effect of interactivity on users' perception, so we cannot confirm our **H3**. Here, despite the predominant quick process based mostly on hunches that characterize more intuitive decision-makers [24], we believe that looking at verbatim excerpts from other users' reviews may also be of benefit to them, to corroborate whether their hunches are aligned with the system's assertions, although they may not do so as extensively as less intuitive users would do.

Additionally, in line with our **H4** and results reported by [26], we observed that social awareness might moderate the effect of interactivity on explanation persuasiveness. Here, results suggest that participants with a higher disposition to listen and take into account others' opinions, tend to perceive higher interactive explanations as more persuasive, which seems a consequence of the possibility to read reports of personal experiences by customers, who have already made use of the recommended items. This represents a potential advantage in the evaluation of experience goods like hotels, which is characterized by a greater reliance on word-of-mouth [29,42].

In regard to our **RQ2** and **RQ3**, and in line with **H5a**, our observations suggest that intuitive decision style might mediate the effect of presentation on explanation satisfaction, independent of interactivity. Particularly, explanatory arguments presented as a bar chart seemed to be perceived as more satisfactory to more intuitive users, than the presentation using a table or only text, presumably due to their greater immediacy [7], thus facilitating the rapid decision-making process that characterizes more intuitive users. However, and contrary to our expectations, we cannot conclude that users with more visualization familiarity will perceive the bar chart explanations better than the text-based ones (**H5b**). One possible reason could be that a text-based format makes it easier to visualize argumentative components as rebuttal and refutation, which could lead to a higher acceptance of an argument, as advocated by argumentation theory ([23]), but could hardly be expressed through graph-based formats.

Additionally, although users with lower visualization familiarity tended to use the interaction options more, we cannot confirm our hypothesis that those users would perceive graphic-based explanations (i.e. bar chart) better when more interactive options are offered, (**H6**). Actually, we found that users with more experience with data visualization reported a more positive perception for explanations with higher interactivity, independent of presentation style. We believe this is not due to difficulties understanding the explanations (as we thought would be the case for users with less visualization familiarity), but because higher interactivity facilitated a structured navigation and more appealing display of the data, which would not be as easy to process or useful if presented on a single static explanation.

Overall, we observed a main effect of rational decision-making style and social awareness in the perception of the system and all the proposed explanations. This suggests that review-based explanations seem to benefit more the users who tend to evaluate information thoroughly and take into account the opinions of others when making decisions, compared to users who use a more shallow information-seeking process.

Interactivity and Transparency Perception. Despite the main effect of interactivity on the overall perception of the system and its explanations, the mean perception of system transparency (user understands why items were recommended) is only slightly higher for the interactivity *high* condition than for the *low* condition. We believe that the reason might be two-fold: 1) Walton's [62] suggests to include an explicit mechanism to confirm effective understanding by the user, so that if this has not yet been achieved, the iterative cycle of user questions and system responses may continue. In consequence, we believe that a more flexible approach in which the user could, for example, write their own questions, rather than the bounded link-based options, might contribute in this regard. And 2) users may be also interested in understanding the reasons why the hotel x is better than hotel y. This would not only be in line with the view of authors who claim that the *why-questions* ask for a contrastive explanation ("why P rather than Q?") [27,34,40], but also concurs with some participants' suggestions, that options for comparison would be very useful.

Use of Interaction Options. We observed that almost half of participants under the condition interactivity "high" actually used the interaction options, although participants were not explicitly instructed to use them, so it can reasonably be inferred that their use was mainly voluntary. It is critical, however, that these options are named appropriately, indicating clearly their destinations (as stated by [20] guidelines), to increase the probability of their use, as evidenced by the lack of use of the option to read reviews excerpts in the *text* condition (Fig. 3e).

Additionally, some of the users assigned to the *low* interactivity condition pointed to 1) the lack of access to additional information in connection to the explanations (particularly customer reviews) as a disadvantage, with about a quarter of those participants writing suggestions on the subject, e.g. "I would prefer to read the actual reviews and understand why ratings were what they were", or 2) insufficiency of aggregated percentages of positive and negative opinions to adequately explain recommendations, e.g. "I feel they maybe could have a lot more information more on SPECIFICALLY what they say about the room instead of just an overall aggregation". In this regard, it is important to note though, that participants of all conditions had access to the full hotel reviews (they were included in the general view of each hotel).

Practical Implications. Our approach was specifically tested in hotels domain, however, since it allows users to navigate from aggregated accounts of other users' opinions to detailed extracts of individual reviews, we believe it might generalize adequately to domains that involve the evaluation of experience goods [43], and where the search for information is characterized by a greater reliance on word-of-mouth [29,42] for example restaurants, movies or books. Additionally, our findings lead to the following practical implications:

- Providing interactive explanations resembling an argumentative communication between system and user could contribute to a better perception of the system, which could be done using web navigation options, e.g. links or buttons, indicating a *why* or *what* questions to be answered by the system.
- Presenting both aggregated opinion statistics and excerpts of comments filtered by feature, as part of an interactive explanation, is a beneficial way to provide explanations sufficient in content, while avoiding overwhelming users with irrelevant data in a single step or screen.
- Given the practical difficulty of detecting user characteristics (e.g., decision-making style or visualization familiarity) by the system, we suggest interactive options to be considered, not only to provide in-depth arguments or to detect the relevance of features to the user, but also to modify the presentation style of argument components.

7 Conclusions and Future Work

In this paper, we have presented a scheme for explanations as interactive argumentation in review-based RS, inspired by dialogue explanation models and formal argument schemes, that allows users to navigate from aggregated accounts of other users' opinions to detailed extracts of individual reviews, in order to facilitate a better understanding of the claims made by the RS. We tested an implementation of the proposed scheme in the hotels domain, and found that more interactive explanations contributed to a more positive perception of effectiveness and trust in the system. We also found that individual differences in terms of user characteristics (e.g. decision-making style, social awareness and visualization familiarity) may lead to differences in the perception of the proposed implementation.

While our proposal suggests a first step towards an effective implementation of interactive explanations for review-based RS, we acknowledge that a major limitation of our approach is that the questions that users can ask are limited to a given set of pre-defined statements. To mitigate such a limitation, we will extend, in future work, our proposed scheme to support a wider range of questions, which can be asked by the user even in their own words. To this end, we plan to leverage advances of conversational agents (i.e. chatbots), natural language processing and natural language generation techniques, such as question answering and automatic summarization, to enhance the implementation proposed in this paper.

Likewise, we plan in the future to investigate the effect of contrastive dialog-based explanations of the type "Why P rather than not-P?". This type of explanation can be leveraged to enable users to influence the recommendation process itself, e.g. requesting for a more refined set of recommendations that better suit their preferences, based on an explanatory contrast between options. This might result in greater satisfaction with the overall system, as has been proven with interactive RS in the past, but this time from the explanations as such.

Acknowledgements. This work was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 2018, p. 1–18 (2018)
2. Arioua, A., Croitoru, M.: Formalizing explanatory dialogues. In: Beierle, C., Dekhtyar, A. (eds.) SUM 2015. LNCS (LNAI), vol. 9310, pp. 282–297. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23540-0_19
3. Bader, R., Woerndl, W., Karitnig, A., Leitner, G.: Designing an explanation interface for proactive recommendations in automotive scenarios. In: Ardissono, L., Kuflik, T. (eds.) UMAP 2011. LNCS, vol. 7138, pp. 92–104. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28509-7_10
4. Bauman, K., Liu, B., Tuzhilin, A.: Aspect based recommendations: recommending items with the most valuable aspects based on user reviews. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 717–725 (2017)
5. Bentahar, J., Moulin, B., Belanger, M.: A taxonomy of argumentation models used for knowledge representation. Artif. Intell. Rev. **33**(3), 211–259 (2010)
6. Berkovsky, S., Taib, R., Conway, D.: How to recommend?: user trust factors in movie recommender systems. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, pp. 287–300 (2017)
7. Blair, J.A.: The possibility and actuality of visual arguments. In: Tindale, C. (eds.) Groundwork in the Theory of Argumentation, vol. 21, pp. 205–223 (2012)
8. Carenini, G., Cheung, J.C.K., Pauls, A.: Multi document summarization of evaluative text. Comput. Intell. **29**, 545–574 (2013)
9. Carenini, G., Moore, J.D.: Generating and evaluating evaluative arguments. Artif. Intell. **170**, 925–952 (2006)
10. Casel: 2013 casel guide: Effective social and emotional learning programs - preschool and elementary school edition, collaborative for academic social and emotional learning (2013)
11. Chen, C., Zhang, M., Liu, Y., Ma., S.: Neural attentional rating regression with review-level explanations. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web, pp. 1583–1592. International World Wide Web Conferences Steering Committee (2018)
12. Chen, L., Pu, P.: Critiquing-based recommenders: survey and emerging trends **22**(1–2), 3085–3094 (2014)
13. Cheng, H.F., et al.: Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)

14. Costa, F., Ouyang, S., Dolog, P., Lawlor, A.: Automatic generation of natural language explanations. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, pp. 57:1–57:2 (2018)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019)
16. Dong, R., O'Mahony, M.P., Smyth, B.: Further experiments in opinionated product recommendation. In: Lamontagne, L., Plaza, E. (eds.) ICCBR 2014. LNCS (LNAI), vol. 8765, pp. 110–124. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11209-1_9
17. Donkers, T., Kleemann, T., Ziegler, J.: Explaining recommendations by means of aspect-based transparent memories. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 166–176 (2020)
18. Donkers, T., Ziegler, J.: Leveraging arguments in user reviews for generating and explaining recommendations. Datenbank-Spektrum **20**(2), 181–187 (2020)
19. Driver, M.J., Brousseau, K.E., Hunsaker, P.L.: The dynamic decision maker (1990)
20. Farkas, D.K., Farkas, J.B.: Guidelines for designing web navigation. Tech. Commun. **47**(3), 341–358 (2000)
21. Faul, F., Erdfelder, E., Lang, A.G., Buchner, A.: G*power 3: a flexible statistical power analysis for the social, behavioral, and biomedical sciences. Behav. Res. Methods **39**, 175–191 (2007)
22. Friedrich, G., Zanker, M.: A taxonomy for generating explanations in recommender systems. AI Mag. **32**(3), 90–98 (2011)
23. Habernal, I., Gurevych, I.: Argumentation mining in user-generated web discourse. Comput. Linguist. **43**(1), 125–179 (2017)
24. Hamilton, K., Shih, S.I., Mohammed, S.: The development and validation of the rational and intuitive decision styles scale. J. Pers. Assess. **98**(5), 523–535 (2016)
25. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 241–250. ACM (2000)
26. Hernandez-Bocanegra, D.C., Donkers, T., Ziegler, J.: Effects of argumentative explanation types on the perception of review-based recommendations. In: Adaptation and Personalization (UMAP 2020 Adjunct) (2020)
27. Hilton, D.J.: Conversational processes and causal explanation. Physcol. Bull. **107**(1), 65–81 (1990)
28. Kirby, J.R., Moore, P.J., Schofield, N.J.: Verbal and visual learning styles. Contemp. Educ. Psychol. **12**(2), 169–184 (1988)
29. Klein, L.: Evaluating the potential of interactive media through a new lens: search versus experience goods. J. Bus. Res. **41**, 195–203 (1998)
30. Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. In: User Modeling and User-Adapted Interaction, pp. 441–504 (2012)
31. Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., Getoor, L.: Personalized explanations for hybrid recommender systems. In: Proceedings of 24th International Conference on Intelligent User Interfaces (IUI 19), pp. 379–390. ACM (2019)
32. Krause, J., Perer, A., Ng, K.: Interacting with predictions: visual inspection of black-box machine learning models. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5686–5697 (2016)
33. Lamche, B., Adigüzel, U., Wörndl, W.: Interactive explanations in mobile shopping recommender systems. In: Proceedings of the 4th International Workshop on

Personalization Approaches in Learning Environments (PALE 2014), held in conjunction with the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP 2014), pp. 92–104 (2012)

34. Lipton, P.: Contrastive explanation. Royal Inst. Philos. Suppl. **27**, 247–266 (1990)
35. Liu, Y., Shrum, L.J.: What is interactivity and is it always such a good thing? implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. J. Advert. **31**(4), 53–64 (2002)
36. Loepp, B., Herrmanny, K., Ziegler, J.: Blended recommending: integrating interactive information filtering and algorithmic recommender techniques. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 2015, pp. 975–984 (2015)
37. Loepp, B., Hussein, T., Ziegler, J.: Choice-based preference elicitation for collaborative filtering recommender systems. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI 2014, pp. 3085–3094 (2014)
38. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: A grounded interaction protocol for explainable artificial intelligence. In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019, pp. 1–9 (2019)
39. McKnight, D.H., Choudhury, V., Kacmar, C.: Developing and validating trust measures for e-commerce: an integrative typology. Inf. Syst. Res. **13**, 334–359 (2002)
40. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2018)
41. Muhammad, K.I., Lawlor, A., Smyth, B.: A live-user study of opinionated explanations for recommender systems. In: Intelligent User Interfaces (IUI 2016), vol. 2, pp. 256–260 (2016)
42. Nelson, P.J.: Consumer Information and Advertising. In: Galatin, M., Leiter, R.D. (eds.) Economics of Information. Social Dimensions of Economics, vol. 3. Springer, Dordrecht (1981). https://doi.org/10.1007/978-94-009-8168-3_5
43. Nelson, P.: Information and consumer behavior. J. Polit. Econ. **78**(2), 311–329 (1970)
44. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Model User Adap. **27**, 393–444 (2017)
45. Perugini, M., Gallucci, M., Costantini, G.: A practical primer to power analysis for simple experimental designs. Int. Rev. Soc. Psychol. **31**(1)(20), 1–23 (2018). https://doi.org/10.5334/irsp.181
46. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems - RecSys 2011, pp. 157–164 (2011)
47. Rago, A., Cocarascu, O., Bechlivanidis, C., Toni, F.: Argumentation as a framework for interactive explanations for recommendations. In: Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning, pp. 805–815 (2020)
48. Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: Methods and metrics for cold-start recommendations. In: Proceedings of SIGIR 2002, pp. 253–260 (2002)
49. Schnotz, W.: Integrated model of text and picture comprehension. In: The Cambridge Handbook of Multimedia Learning, 2nd ed., pp. 72–103 (2014)
50. Sniezek, J.A., Buckley, T.: Cueing and cognitive conflict in judge advisor decision making. Organ. Behav. Hum. Decis. Process. **62**(2), 159–174 (1995)

51. Sokol, K., Flach, P.: LIMEtree: interactively customisable explanations based on local surrogate multi-output regression trees. arXiv preprint arXiv:2005.01427 (2020)
52. Sokol, K., Flach, P.: One explanation does not fit all: the promise of interactive explanations for machine learning transparency **34**(2), 235–250 (2020)
53. Song, J.H., Zinkhan, G.M.: Determinants of perceived web site interactivity. J. Mark. **72**(2), 99–113 (2008)
54. Steuer, J.: Defining virtual reality: dimensions determining telepresence. J. Commun. **42**(4), 73–93 (1992)
55. Tintarev, N.: Explanations of recommendations. In: Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys 2007, pp. 203–206 (2007)
56. Tintarev, N., Masthoff, J.: Evaluating the effectiveness of explanations for recommender systems. User Model. User Adapt. Interact. **22**, 399–439 (2012)
57. Toulmin, S.E.: The uses of argument (1958)
58. Vig, J., Sen, S., Riedl, J.: Tagsplanations: explaining recommendations using tags. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, pp. 47–56. ACM (2009)
59. Wachsmuth, H., Trenkmann, M., Stein, B., Engels, G., Palakarska, T.: A review corpus for argumentation analysis. In: 15th International Conference on Intelligent Text Processing and Computational Linguistics, pp. 115–127 (2014)
60. Walton, D.: The place of dialogue theory in logic. Comput. Sci. Commun. Stud. **123**, 327–346 (2000)
61. Walton, D.: A new dialectical theory of explanation. Philos. Explor. **7**(1), 71–89 (2004)
62. Walton, D.: A dialogue system specification for explanation. Synthese **182**(3), 349–374 (2011)
63. Walton, D., Krabbe, E.C.W.: Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. State University of New York Press, New York (1995)
64. Wang, N., Wang, H., Jia, Y., Yin, Y.: Explainable recommendation via multi-task learning in opinionated text data. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, pp. 165–174 (2018)
65. Weld, D.S., Bansal, G.: The challenge of crafting intelligible intelligence. Commun. ACM **62**(6), 70–79 (2019)
66. Wu, Y., Ester, M.: Flame: a probabilistic model combining aspect based opinion mining and collaborative filtering. In: Eighth ACM International Conference on Web Search and Data Mining, pp. 153–162. ACM (2015)
67. Xiao, B., Benbasat, I.: Ecommerce product recommendation agents: use, characteristics, and impact. MIS Q. **31**(1), 137–209 (2007)
68. Yaniv, I., Milyavsky, M.: Using advice from multiple sources to revise and improve judgments. Organ. Behav. Hum. Decis. Process. **103**, 104–120 (2007)
69. Zanker, M., Schoberegger, M.: An empirical study on the persuasiveness of fact-based explanations for recommender systems. In: Joint Workshop on Interfaces and Human Decision Making in Recommender Systems, pp. 33–36 (2014)
70. Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma., S.: Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 83–92 (2014)

Paper 4

The following paper is reused from:

- Hernandez-Bocanegra, D.C., & Ziegler, J. (2021). Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *3rd Conference on Conversational User Interfaces (CUI '21)*, 1-11. ACM, New York, NY, USA. doi: https://doi.org/10.1145/3469595.3469596

Conversational review-based explanations for recommender systems: Exploring users' query behavior

Diana C. Hernandez-Bocanegra
University of Duisburg-Essen
Duisburg, Germany
diana.hernandez-bocanegra@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

Providing explanations based on user reviews in recommender systems (RS) can increase users' perception of system transparency. While static explanations are dominant, interactive explanatory approaches have emerged in explainable artificial intelligence (XAI), so that users are more likely to examine system decisions and get more arguments supporting system assertions. However, little attention has been paid to conversational approaches for explanations targeting end users. In this paper we explore how to design a conversational interface to provide explanations in a review-based RS, and present the results of a Wizard of Oz (WoOz) study that provided insights into the type of questions users might ask in such a context, as well as their perception of a system simulating such a dialog. Consequently, we propose a dialog management policy and user intents for explainable review-based RS, taking as an example the hotels domain.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Human-centered computing → User studies**; **Natural language interfaces**.

KEYWORDS

Recommender systems, explanations, argumentation, conversational agent, user study

ACM Reference Format:
Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3469595.3469596

1 INTRODUCTION

Customer reviews have been increasingly used for explaining decisions made by recommender systems (RS), due to their wealth of detailed information on positive and negative aspects of items, which cannot be obtained directly from ratings. Although review-based explanations can be useful in improving the perception of efficacy and trust in RS, these are almost always presented in a static manner, often as an aggregation of opinions, limiting users in exploring the diverse views and arguments expressed in the reviews. On the other hand, interactive methods may positively influence user perception of RS [22], by allowing the user to request, for example, further elaboration of the claims made by the system. However, explanatory methods that allow users to scrutinize and customize explanations through interaction are largely unexplored, or lack sufficient empirical evidence [56]. Additionally, most interactive approaches in RS and, in a wider scope, in explainable artificial intelligence (XAI), are based on point and click options. However, recent developments in natural language processing (NLP) and natural language generation (NLG) enable a more flexible interaction, where users could indicate, in their own words, their explanation needs.

In particular, we aim to explore the feasibility and implications of using conversational approaches to explanations in review-based RS, and in particular the use of conversational agents (CA), given their ability to enable two-way natural language communication, opening up the range of possible questions a user can ask the system, which could contribute to a better understanding and acceptance of explanations by users, as prescribed by conceptual models of explanation based on dialogue [23, 62]. Although user interfaces inspired by human-to-human conversation have been developed and used for a long time to assist users in a wide range of tasks [46], little is known about how a CA should be conceptualized or designed in the context of XAI, and in particular, in explainable RS. Thus, we aim to explore:

RQ1: How to design a dialog management policy to implement a CA with explanatory purposes in review-based RS?

In this paper, we focus on the analysis of conversation patterns within an explanatory process. In this regard, [43] have drawn attention to the social and communicative aspect of explanation ("someone explains something to someone" [23]) and how an interactive and conversational approach could contribute to increasing user understanding in XAI approaches. While a general theoretical model of explainable recommendations has not yet been established, we propose to analyze explanations through the lens of argumentation theory. A first category of argumentation models seeks to define logical structures containing assertions, supporting evidence, refutations, among others [7]. A second category involves dialectical approaches [63], focusing on the exchange of arguments and supporting (or contradictory) information within a dialogue between two parties.

Thus, our goal is to explore the modeling of explanations in review-based RS as an argumentative dialogue, and how this can be facilitated by a conversational user interface. However, the above

CUI '21, July 27–29, 2021, Bilbao (online), Spain

ACM ISBN 978-1-4503-8998-3/21/07…$15.00
https://doi.org/10.1145/3469595.3469596

requires a close understanding of how a user would formulate questions in this particular setting. Particularly we aim to answer:

RQ2: How do review-based RS users communicate their explanation needs using a CA?

To this end, we conducted a WoOz study [26], taking as an example the hotels domain, since it represents an interesting mix between search goods (with attributes on which complete information can be found before purchase [49]) and experience goods (which cannot be fully known until purchase [49]). Such a product evaluation could benefit from third-party opinions [27, 49], potentially rich in argumentative information that can be used for explanatory purposes in RS. The results of our analysis provided a basis to formulate a dialog management policy for explainable review-based RS, and to draw attention to the challenges involved in implementing such an approach. The contributions of this paper can be summarized as follows:

- We propose a dialog management policy for explanations as conversational argumentation in review-based RS, inspired by dialog models and argument theories.
- We modeled the intents that can be used for the implementation of a CA for explanatory purposes in the hotels domain, based on a WoOz study, and analyzed to what extent follow-up questions were formulated.
- Participants' perception of a simulated system was evaluated in terms of system transparency, trust and effectiveness, as well as satisfaction with the explanation, sufficiency, confidence and persuasiveness.

2 RELATED WORK

2.1 Review-based and argumentative explanations

Explanations can bring several benefits to RS, by increasing users' perception of transparency, effectiveness, and trust [58]. Review-based explanatory RS integrate ratings and reviews to generate both predictions and explanations (e.g. [6, 64, 68]), usually presented as summaries of the positive and negative opinions on different aspects (e.g. [48]). Moreover, exploitation of reviews can facilitate the generation of argumentative explanations, [20], in which system claims (user will find a recommended item useful) are supported by evidence found and consolidated from reviews.

While argumentative approaches have already been applied to explanations, these are mainly based on the static display of the arguments, as in [4, 11, 20, 30, 66], where little can be done to indicate to the system that additional information is still needed to fully understand and accept the explanations. In contrast, interactive and conversational approaches to explanations seek to grant users further control over explanatory components [22, 56], in order to promote a better understanding of the rationale behind system predictions, based on the idea of an exchange of questions and answers between the user and the system, as would occur in a human explanatory conversation [43].

2.2 Conversational explanations

Accordingly, formal explanation dialogues have been conceptually formulated as theoretical support to the design of conversational explanation approaches [2, 14, 39, 54, 60]. Interactive and conversational explanations have been already addressed in the field of explainable artificial intelligence (XAI), although to a much lesser extent compared to static explanations [1], and mostly focused on the influence that features or data points have on machine learning predictions. For example, [56] proposed a system that provides explanations as an interactive dialogue that resembles a natural language conversation supporting why-questions, to facilitate the understanding of machine learning classification outcomes, e.g. the rejection of a credit loan. However, this approach differs from ours in that we use non-discrete and non-categorical sources of information, subjective in nature and unstructured, which are nevertheless rich in arguments that can be used to answer questions of a subjective nature. Similarly, [54] defined a protocol to provide conversational argumentative explanations in RS, however it restricts the possible user interactions to a limited set of possible questions a user may ask, while we explore possibilities for users to express their explanatory needs in their own words. Finally, despite the potential benefit of using dialog models to increase users' understanding of intelligent systems [43, 65], their practical implementation in RS (and in XAI in general) still lacks sufficient empirical evaluation [39, 43, 56], thus, it is still unclear how conversational explanatory interfaces should be conceived and designed, so that they actually improve users' perception of RS.

Consequently, we set out in this paper to explore the design of a dialog management policy for conversational explanations in RS, exploiting the potential benefits of a dialog system (particularly a CA or chatbot), where users can indicate their explanatory needs in their own words, in the form of questions. Our work differs from the traditional approach to CA in the hotel domain, which focuses on processes like customer service and booking assistance [10], and to conversational RS (e.g. [13, 67]) which aim to collect user preferences to generate recommendations through dialog. We aim, on the other hand, to explore the implications and effects of using CA to explain RS rationale, which remains largely unexplored [24]. A model of social explanations for movie recommendations was proposed by [51], in one of the few works on the subject. However, according to their approach, it is the system who leads the conversation, providing justifications for recommendations even when they are not explicitly requested by the user, whereas according to our proposal, the user would have the active role, being enabled to ask the questions that lead to an argumentation by the system.

2.3 Question answering (QA)

Our work is closely related to QA systems, which aim to answer questions posed by users in natural language, by using techniques like information retrieval (IR) or NLP, on various types of web documents or in knowledge bases. While most of QA systems are designed to respond factoid, definition, or list questions by offering excerpts from documents or list of items consistent with user's query, much less work has been devoted to advanced "how-to", "why", evaluative, comparative, and opinion questions [34, 44], that require usually the aggregation and comparison of multiple items over different pieces of information. Lipton [35] defines *explanation* as an answer to a *why-question* , however, other types of questions can also be answered by explanations, i.e. how? what? [43], the latter being one that could be answered with a factoid sentence, for which we aim to support both factoid and advanced question types. Additionally, and in contrast to the common QA approach where the system replies to a series of standalone questions, interactive QA involves a dialog interface enabling related, follow-up and clarification questions [53].

Nevertheless, our approach differs from most QA methods, especially those based on IR, because in our case, responses should not be generated solely on the basis of information sources, but should be consistent and reflect the mechanism used to generate the recommendations. Additionally, to answer complex questions (e.g., "why"), our approach involves a focus on the most relevant aspects for users, to provide concise and relevant statements that aggregate information from different reviews. To this end, our approach relies on the user profile inferred by the RS algorithm, especially when no explicit features are addressed in users' questions. On the other hand, implicit user preferences are not taken into account in most QA approaches, which stems from their use of IR methods, where the relevance of a document is estimated based on how much its content is related to the query [45]. Additionally, we propose to follow an argumentative explanation structure to generate responses, which could improve users' perception of RS, as evidenced using the interactive, although not fully conversational approach proposed by[22]. Although argumentation has already been exploited in QA [47], it has been mostly used to extract high quality answers by means of argument mining, whereas very few approaches exploit argumentation as a way of presenting explanations in response to user queries [3].

2.4 Users' utterances on explanation needs

The design of adequate conversational explanations requires a proper understanding of possible user requests [33], which may vary according to the type of application, the context and user characteristics. [32] collected a dataset consisting of written conversations between humans with a movie recommendation goal, however, no explanatory inquiries like "Why do you recommend X?" are addressed . Furthermore, [8] collected a QA dataset for several domains (including hotels), which can be used to generate answers not limited to factoid questions, but also to subjective ones (e.g. "How is it the location?"). However, questions and answers only address one item at a time, leaving out comparison queries; moreover, the dataset is not oriented as such to an explanatory dialogue. On the other hand, [33] noted that a question-oriented framework offers a feasible way to conceive interactive explanations, and proposed a XAI question bank, consisting of inquiries that users might typically ask about AI algorithms. However, as it is the case for most XAI interactive approaches, this question bank was intended for explanation needs of users with expert knowledge in AI, whereas no similar question bank definition has been developed, to our knowledge, for end users and, in particular, for RS. Consequently, we conducted a user study using the WoOz [26] method, to capture the possible questions users would ask in the context of RS explanations, particularly in the hotel domain.

2.5 WoOz paradigm

WoOz studies allow for the incremental design of conversational interfaces, and involve the simulation of a human-machine interaction, in which a member of the research team (the *wizard*) simulates the response actions of the system, through a computer-mediated interface, a technique that has been widely adopted for HCI prototyping [15, 40]. The use of this type of technique allows to validate how users would interact with a conversational interface, and to evaluate the feasibility of dialog based systems that have not yet been fully implemented, as was done for example by [53] to design an open domain interactive QA system, or by [5] for the design of a conversational framework to support recipe recommendation.

3 EXPLANATIONS AS CONVERSATION

Our proposal is based on previous work reported in [22], where the effect of different levels of interactivity for accessing explanatory information was tested, without considering a CA perspective as such. Such approach was inspired by dialog-based explanation models [39, 61, 62] and the argumentative scheme by [19], and regards an explanation as an interactive argumentation, that is, an explanation consisting of a cyclic sequence of *argumentation attempts* made by the system in response to *argument requests* made by the user, as a way to challenge or critique system argumentation, or to inquire for further arguments, using why, what, or how-questions (Figure 1). Argumentation attempts include premises (a general reason to accept a claim that a recommended item is worthy to be chosen) and backing (specific information or additional evidence to support the claim, e.g. percentage of positive opinions about an aspect), among others.

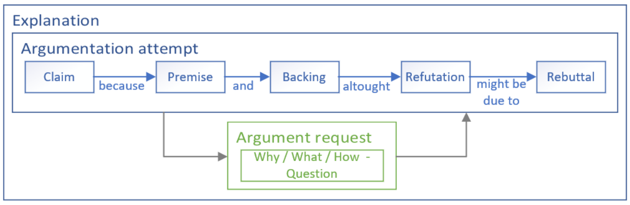


Figure 1: Simplified scheme for explanations as interactive argumentation in review-based RS [22].

Despite the positive perception by users of a system that implemented such a scheme, its components were not directly derived from observed natural human conversation, leading to the following constraints: 1) it only offers answers to a limited set of questions, 2) it does not consider comparative questions, e.g. "why is X better than Y?", 3) nor factoid questions, e.g. "does this hotel include breakfast?", 4) nor questions regarding users' own profile, or algorithm details. In consequence, we extended this scheme to support a wider range of questions that could be written by users in their own words, and used it as basis for the valid moves of the wizard in our study (Figure 2). Refutation and rebuttal components from proposal in [22] were left out from current proposal, for the sake of brevity of responses by the system (following guidelines by [46]), and will be evaluated in future work.

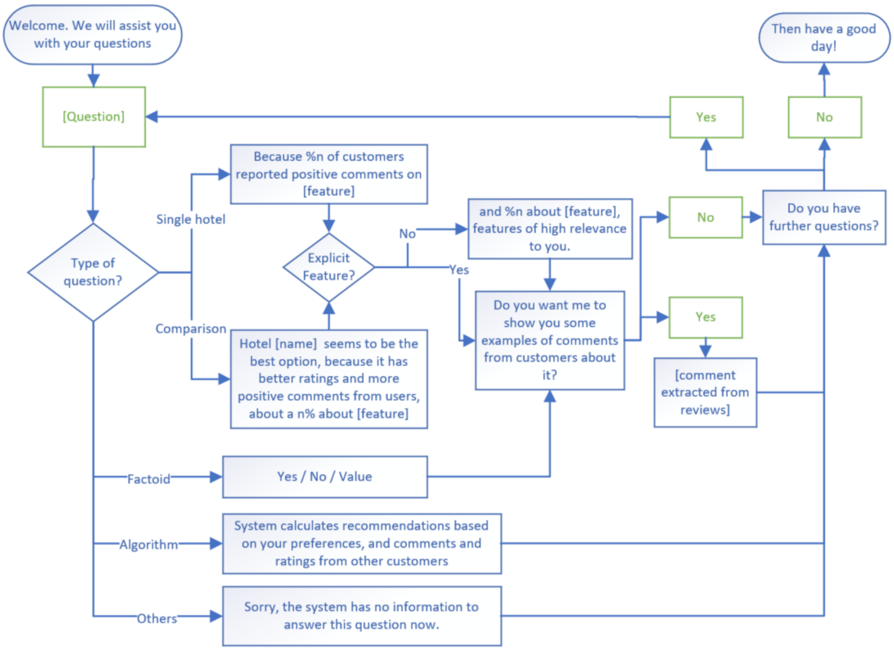

Welcome. We will assist you with your questions
[Question]
Type of question?
Single hotel
Comparison
Because %n of customers reported positive comments on [feature]
Explicit Feature?
Hotel [name] seems to be the best option, because it has better ratings and more positive comments from users, about a n% about [feature]
No
Yes
and %n about [feature], features of high relevance to you.
Do you want me to show you some examples of comments from customers about it?
No
Yes
[comment extracted from reviews]
Do you have further questions?
Yes
No
Then have a good day!
Factoid
Yes / No / Value
Algorithm
System calculates recommendations based on your preferences, and comments and ratings from other customers
Others
Sorry, the system has no information to answer this question now.

Figure 2: Scheme for conversational explanations used in WoOz experiment. Blue boxes represent utterances by the system, green boxes the utterances by users.

4 USER STUDY

We conducted a WoOz study to explore how users would express their explanatory needs, to a CA in a review-based RS, with hotels as an example domain. All subjects were assigned to the same experimental condition, and were instructed to interact with the RS, and to write their questions about the reasons for the recommendations, which were replied by the wizard (played by our main researcher). We used the scheme described in the previous section (Fig 2), as the guideline for the wizard, aiming to portray a structured conversation similar across participants. Particularly, we hypothesize that users will ask questions of the types *why?*, *how?*, and *what?*, as well as factoid, comparative, and evaluative questions, at the feature level as well as at the general level. Further details about the study are described below.

4.1 Participants

We recruited 20 participants (10 female, mean age 34.65 and range between 20 and 69) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 98%, and a number of HITs approved greater than 500. Participants were informed in consent form and instructions about payment rejection (if no effective interaction with the system) which could be checked using system logs, and responses to validation questions in questionnaires (e.g. "recommendations were based on: Opinions of celebrities, True/False", "The purpose of this question is to check attentiveness, please mark Disagree"). We discarded participants with less than 5 (out of 7) correct answers, or no effective interaction with the wizard. The responses of 12 of the 32 initial participants were then discarded for a final sample of 20 subjects. Participants were rewarded with $2 plus a bonus up to $0.40 depending on the quality of their response to the question "Why did you choose this hotel?" set at the end of the survey, aiming to achieve a more motivated choice by the participants, and to encourage an effective interaction with the system. Time devoted to the task by participants (in minutes): M=12.99, SD= 2.24.

4.2 Procedure

We informed participants that a list of hotels reflecting the results of a hypothetical search and within the same price range would be presented (i.e no filters to search hotels were offered), and that they could consult the general hotel information (photos, reviews, etc., by clicking on "Info Hotel"), but also freely enter any question of interest about one or more hotels in the chat box located on the right of the hotel list. We underlined that the chat box was designed to explain the reasons for the recommendations, in order to prevent the user from asking questions about other processes, such as booking assistance. We presented a cover story, to establish a common starting point in terms of travel motivation, asking participants to imagine the planning of a vacation trip, as in pre-COVID19 times, and that they had to decide which hotel to stay at. We requested the 5 most important hotel aspects to the participant, ranked in order of importance, to calculate personalized recommendations. We then presented the system showing a list of 6 recommended hotels (sorted by predicted recommendation score) and the "chatbox" (Figure 3). A debrief was provided at the end, indicating the main objective of the study.

4.3 Ethic concerns

The WoOz technique relies on deception: participants are supposed to believe they are interacting with a system, so researchers can have a better perception of what users would do when interacting with a real machine. Such a set up raises some ethical concerns given the necessary deception [15]. Following guidance from [17, 50, 57], we took the following actions to mitigate negative effects due to the study deception:

- We avoided an explicit mention of a "full automated" system or chatbot, instead we referred to a "chat box", where they could type their questions.
- We disclosed in the debrief that the responses were written by a human, and that the participants could request for the withdrawal of their responses in case they consider that the procedure went against their expectations, with payment being processed anyway.
- The main study researcher played the wizard, following a pre-established dialog flow, to avoid statements out of project scope that could harm or make participants uncomfortable.

4.4 Dataset and implemented system

Dataset. We used the ArguAna dataset [59], (hotel reviews and ratings from TripAdvisor; sentiment and explicit features annotated sentence wise), and the aspect annotation done by [22], in order to provide aspect based arguments.

Explainable RS. We used the review-based RS developed by [22], which implements the matrix factorization model proposed by [68], in combination with sentiment-based aspect detection methods, using the state of art NLP model BERT [16].

Conversational interface. We used Flask-SocketIO, a Socket.IO integration for Flask applications [18], to allow communication between participants and the wizard. Figure 3 depicts the interface presented to participants

Support system. To obtain the desired benefits, the wizard had to produce responses as fast and consistently as possible, so that users still feel they were interacting with a machine. This can only be achieved if the wizard uses a suitable support system [15], that provides, beyond canned sentences, appropriate answers consistent with participants preferences and the information they can obtain in their own system view. Thus, we added a module to the RS to quickly generate the answers, so the wizard could copy and paste them in the conversational interface.

Personalization mechanism: To reduce implications of the *cold start* problem [55] (system does not have enough information about the user to generate an adequate profile and thus, personalized recommendations), participants were asked for their aspects of most importance, and the RS selected the user with the highest preference similarity within the rating matrix of the RS algorithm to generate predictions.

4.5 Questionnaires

System perception. We evaluated system perception based on explanatory aims defined by Tintarev [58]. We focused on the subset effectiveness and trust, for which a significant effect of interactive options to explain was found in [22], and on transparency, and on transparency, for which an effect of conversational features is expected, as predicted by the dialogue models of explanation [62]. We utilized items for transparency [52] (user understands why items were recommended), effectiveness [28] (internal reliability Cronbach's α = 0.94, system is useful and helps the user to make better choices), and trust [42] (α = 0.92, user trusts system recommendations).

Explanations perception. We used single items from [29], which involve aspects related to explanations rather than the overall system: explanation confidence (user is confident that she/he would like the recommended item), explanation satisfaction (user would enjoy a system if recommendations are presented this way), and persuasiveness (explanations are convincing), and from [22] for sufficiency (explanations provided are sufficient to make a decision). All items were measured with a 1-5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

4.6 Data Analysis

We first manually classified utterances into categories: questions and no-questions, the latter including e.g. greetings or gratitude statements. We categorized every question according to the dimensions: *scope*, *comparison*, *assessment* and *detail*, following an inductive category formation [41], i.e. we started with one category and benchmarked each question against the criteria of the category. Following that, we either classified the question into the existing category or created a new one. This step involved two independent annotators, who came to a Cohen's kappa = 0.91, almost perfect agreement intercoder reliability [31].

We checked whether questions were standalone questions, or follow-up questions, validating the presence of anaphoras ("a linguistic form whose full meaning can only be recovered by reference to the context") and ellipsis ("an omission of part of the sentence, resulting in a sentence with no verbal phrase") [53]. We used criteria from [53] and [9] to classify anaphoras (pronoun or possessive adjective, and noun phrase anaphora), and ellipsis.

Finally, we evaluated questions according to the feasibility of their automated response, and classified them according to possible methods that could be used to do so.

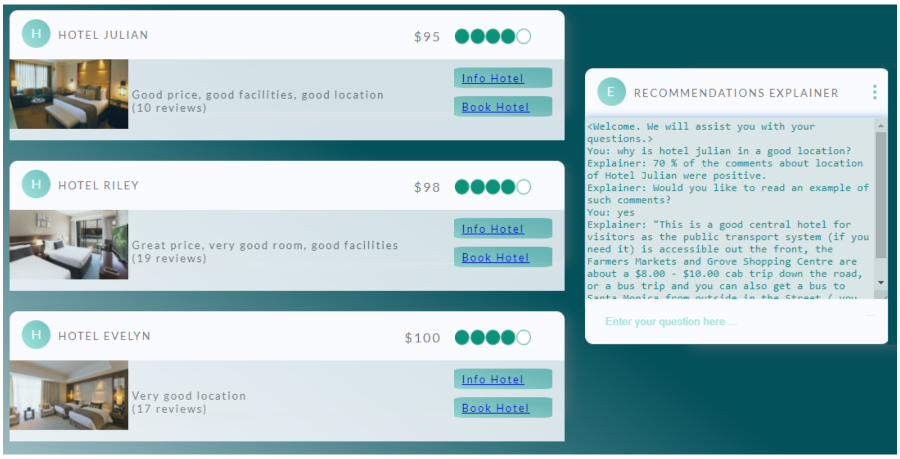
5 RESULTS

We collected a total of 20 dialogues and 105 utterances (M=5.20 utterances per participant, SD=2.48). 81 of the utterances were questions (M=4.05 questions per participant, SD=2.14). The average question length is 46.70 characters (SD=30.86). We observed that the conversations adhered to the explanatory objective, and not to other purposes, such as, booking process.

5.1 Intents and entities

We identified that users' intents could be classified into two main types: *domain-related* intents (regarding hotels and their features), and *system-related* intents (regarding the algorithm, or the system input). In turn, domain-related intents could be categorized according to the following dimensions:

- *Scope*: Whether the question refers to a single item (*single*), a limited list of items (*tuple*), or to no particular item (*indefinite*).



Good price, good facilities, good location
(10 reviews)

Great price, very good room, good facilities
(19 reviews)

Very good location
(17 reviews)

```
<Welcome. We will assist you with your
questions.>
You: why is hotel julian in a good location?
Explainer: 70 % of the comments about location
of Hotel Julian were positive.
Explainer: Would you like to read an example of
such comments?
You: yes
Explainer: "This is a good central hotel for
visitors as the public transport system (if you
need it) is accessible out the front, the
Farmers Markets and Grove Shopping Centre are
about a $8.00 - $10.00 cab trip down the road,
or a bus trip and you can also get a bus to
Santa Monica from outside in the Street ( you
```

Figure 3: User interface presented to participants in WoOz study.

- *Comparison*: Whether the question is (*comparative*) or not (*non-comparative*). We adopt the comparative sentence definition by [25] "expresses a relation based on similarities or differences of more than one object", including superlatives and relations like "greater" or "less than".
- *Assessment*: Whether the question refers to the existence or characteristics of item features (*factoid*), to a subjective assessment of the item or its features (*evaluation*), or to system reasons to recommend an item (*why-recommended*).
- *Detail*: Whether the question inquires for an specific aspect or feature (*aspect*), or for the overall item (*overall*).

Consequently, the intent of a single domain question could be defined as a combination of the 4 dimensions. Table 1 contains examples for every dimension / value, Figure 4 depicts the distribution of questions regarding every dimension, and Table 2 contains examples of intents, and their frequency in the collected utterances. It is important to note that all but one of the questions could be correctly classified as system-related intent, namely: "why are there so few reviews?".

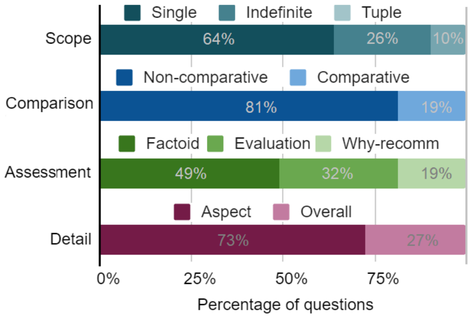


Figure 4: Distribution of questions according to each dimension of domain-related intents.

All questions of domain intent regarded the entities: *hotel* and *hotel feature*.

5.2 Follow-up questions

Figure 5 shows the distribution of standalone and follow-up questions. A special case are inquiries that could work as both types. Such is the case for comparative questions under the value "Indefinite" of dimension *scope*, which may refer to the best among all possible options (e.g. "which is the best hotel?") or, if a subset of options was previously discussed, as a follow-up, (e.g. "I am choosing between the Riley and the Evelyn. Which is the best hotel overall?").

Additionally, Figure 5b shows the distribution of follow-up question types: pronoun or possessive adjective anaphoras (e.g. " I'm looking for facts about current internet service - is **it** unchanged or upgraded?"), noun phrase anaphora (e.g. "When was the last time **the Hotel** underwent a remodel?), and ellipsis (e.g. "what are the checking in times for hotel owen? **and hotel evelyn?**"). We noted that pronouns and noun phrases in anaphoras referred only to hotels names or hotel features.

Moreover, 11% of utterances contained non-question sentences aiming to establish a context for a subsequent question, e.g. "I like the ambiance of the Hotel Evelyn, how were the reviews for that?". Finally, only 2.4% of utterances contained more than one question.

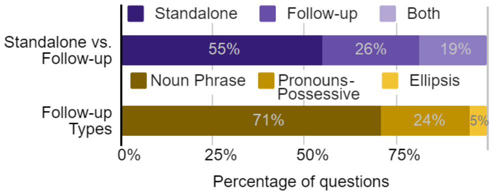


Figure 5: Distribution of follow-up questions.

5.3 Methods for generating answers

The number of questions that could be answered with different types of methods is shown in Figure 6. Some could be replied by

Table 1: Example of domain-related intents classified by dimension.

Dimension	Value	Question is about	Example
Scope	Single	A specific item	How is the food at **Hotel Evelyn**?
	Tuple	Two or more items	Are either hotel **owen** or **evelyn** near station?
	Indefinite	No specific item (s)	Which hotel has the best views?
Comparison	Comparative	Relation of similarities or differences of more than one object.	what is **difference between** hotel evelyn and hotel james? Which hotel has the **best** views?
	Non-comp	No comparison	How close is Hotel Owen to the subway?
Assessment	Factoid	Facts, item having features or not	**Does** Hotel Owen have TV service?
	Evaluation	How hotel is evaluated (subjectively)	**How** is the food at Hotel Evelyn?
	.	Which hotel/feature is better/best	**Which** hotel has the **best** view?
	Why-recomm	Reasons of recommendations	**Why** is Hotel Julian my top recommendation?
Detail	Aspect	A specific aspect or feature	Why is it Hotel Julian in a good **location**?
	Overall	No specific aspect or feature	How good is hotel Julian?

Table 2: Most frequent domain intents (combination of dimensions values) sorted by number of questions per intent (desc.)

Scope	Comparison	Assessment	Detail	Example	# Qs	Type of initial response
Single	Non-comp	Factoid	Aspect	Does Hotel Julian have a pool?	29	Y/N or value
Single	Non-comp	Why-recomm	Overall	Why is Hotel Julian my top recommendation?	14	Because [Argument backing]
Single	Non-comp	Evaluation	Aspect	How is the food at Hotel Evelyn?	8	[Argument claim], because [Argument backing]
Indefinite	Comparative	Evaluation	Aspect	Which hotel has the best customer service?	7	Hotel X, because [Argument backing]
Indefinite	Non-comp	Factoid	Aspect	Do any of the hotels provide free breakfast?	6	Y/N or value
Tuple	Non-comp	Factoid	Aspect	what are the checking in times for hotel owen and hotel evelyn?	4	Y/N or value
Indefinite	Comparative	Evaluation	Overall	Which hotel has the best reviews?	4	Hotel X, because [Argument backing]
Indefinite	Non-comp	Evaluation	Aspect	what rooms would be good for parents with children?	3	Hotel X, because [Argument backing]
Tuple	Comparative	Evaluation	Overall	What is difference between hotel evelyn and hotel james?	2	Hotel X has better comments on [feature x] and [feature y].

using several methods, e.g. "How close is Hotel Julian to the city centre?" could be replied both using hotel metadata, or a QA method to retrieve answers from users comments. Additionally, according to our proposed scheme, a question like "How is the food at Hotel Evelyn?" could be replied by presenting an aggregation of opinions, and by providing an example of such opinions extracted from reviews. Finally, 17% of the questions could not be directly replied to by any of our contemplated methods, e.g. "how was the price of the hotel decided?", given that price is not decided directly by the RS. Although we intended to provide approximate answers to questions such as "Has Hotel Evelyn made any upgrades to its internet/wi-fi service since some of its reviews were written?", such as "X% of customers reported positive opinions about wi-fi", it may not be enough to satisfy very curious users, as in this case, where we got the counter-response: "That doesn't answer my question".

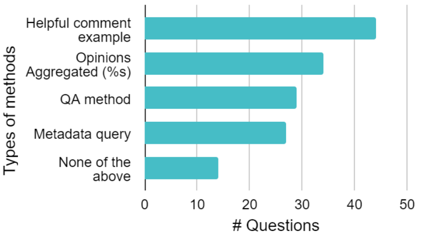


Figure 6: Number of questions that could be responded by different types of methods.

5.4 Perception of system and explanations by users

Figure 7 shows the distribution of users' perception, and the distribution of topics addressed in suggestions and comments provided by participants at the end of the study, in their own words.

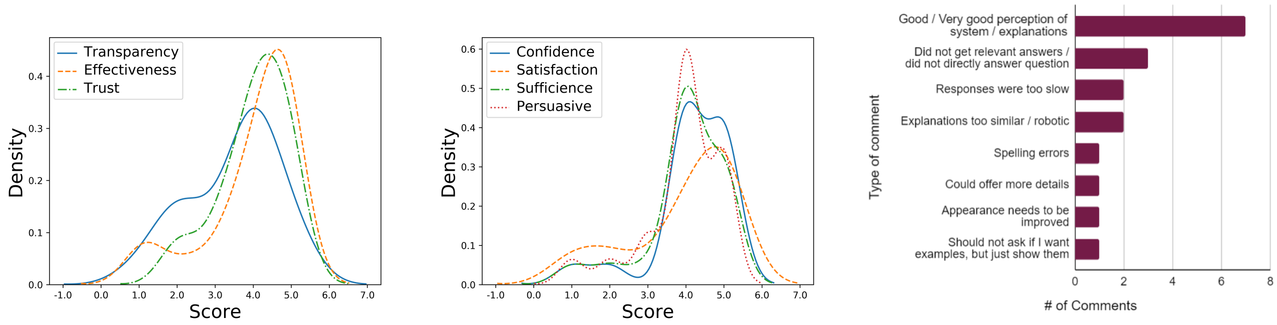


Figure 7: Kernel density estimates of participants' scores for perception of system (left) and explanations (middle); higher score values indicate a more positive perception. Distribution of comments and suggestions from participants (right)

6 DISCUSSION

Suitability of the approach. We consider that our proposed scheme and the WoOz study setup have been useful and effective for our purpose of exploring the use of CA to explainable review-based RS, given the predominantly positive perception of RS and its explanation by participants - especially in terms of effectiveness and trust -, and the observation that collected conversations adhered to the explanatory objective as expected, i.e. no questions regarding other processes were asked, like hotel booking.

Moreover, we observed that users actively expressed their needs for explanation, taking the lead in formulating their own questions (not expecting the system to choose what to explain) and challenging the system's attempts at argumentation when the answers provided did not satisfy their need. We believe that an implementation of our dialog management policy might contribute to a better perception of the RS, in comparison to interfaces providing only static explanations, or interactive but with a very limited set of possible questions to be answered, since 1) it allows for greater active control (voluntary actions that can influence the user experience [36]), which might be beneficial in environments involving information needs and a clear goal in mind [36], and 2) the two-way communication enabled, which might contribute to a better acceptance and understanding of arguments, as predicted by dialog models of explanation [23, 62].

Types of questions. As expected, participants asked both factoid questions and evaluative and why-recommended questions. Although not handled by the method our work is based on (matrix factorization model that integrates reviews [68]), the input from factoid questions could be handled as wish conditions, and lead to changes in recommendations' appearance (highlighting those that match the desired conditions) or to recalculate recommendations' ranking, as is done in critique-based RS. This has been proven to be beneficial to user experience [12, 37, 38], thus we believe it may also be useful to integrate it into our approach, once the factoid response does not remain a flat answer for a single item, but can be applied to the entire set of options, to facilitate comparisons to make a final decision.

Comparing our collected inquiries with the prototypical questions from XAI question bank by [33], we found that their why-questions had a similar objective to the our why-recommended: to ask for reasons why certain predictions have been provided. However, we also observed substantial differences in regard to other types of questions:

- Input questions (e.g. "what kind of data does the system learn from?") were asked only once in our study.
- No questions were asked about output (e.g., "what does the system output mean")
- Neither on performance (e.g. "how accurate or reliable are predictions?").
- We noted that how-questions asked mostly "how the opinions are" rather than asking about the overall logic of the system.
- No "What if?" questions were asked. However, factoid questions might implicitly ask such questions (e.g. "Which hotel has a gym?" could be considered as "what if the system takes into account that 'gym' is an important feature to me?").

These differences could be explained by the context of the task to be performed, and the type of users involved (general public vs. AI experts). However, it was somehow surprising to us that all but one of the questions referred to the system itself, its algorithm, or the input used for predictions. We believe that this may have been due to:

- Users might have perceived that the recommendations matched their preferences and that they had generally positive opinions, i.e., they did not receive very strange recommendations that raised their suspicions.
- Decisions in the chosen domain (hotels) are not as sensitive as in the medical or credit lending domains, where understanding the system logic or input influencing the prediction is critical to the acceptance of the system arguments.
- The perspective and opinion of others might be more relevant than details about their own inferred profile, as reported by [21] for opinionated explanations in a hotel RS, which seems to be the case when evaluating experience goods like hotels, a process characterized by a greater reliance on word-of-mouth [27, 49]. .

In regard to the *scope* dimension, we observed a dominance of single item questions. Although some authors consider that explanations mainly respond to contrast questions ("why P rather than Q?") [23, 35, 43], we observed that the comparative questions with

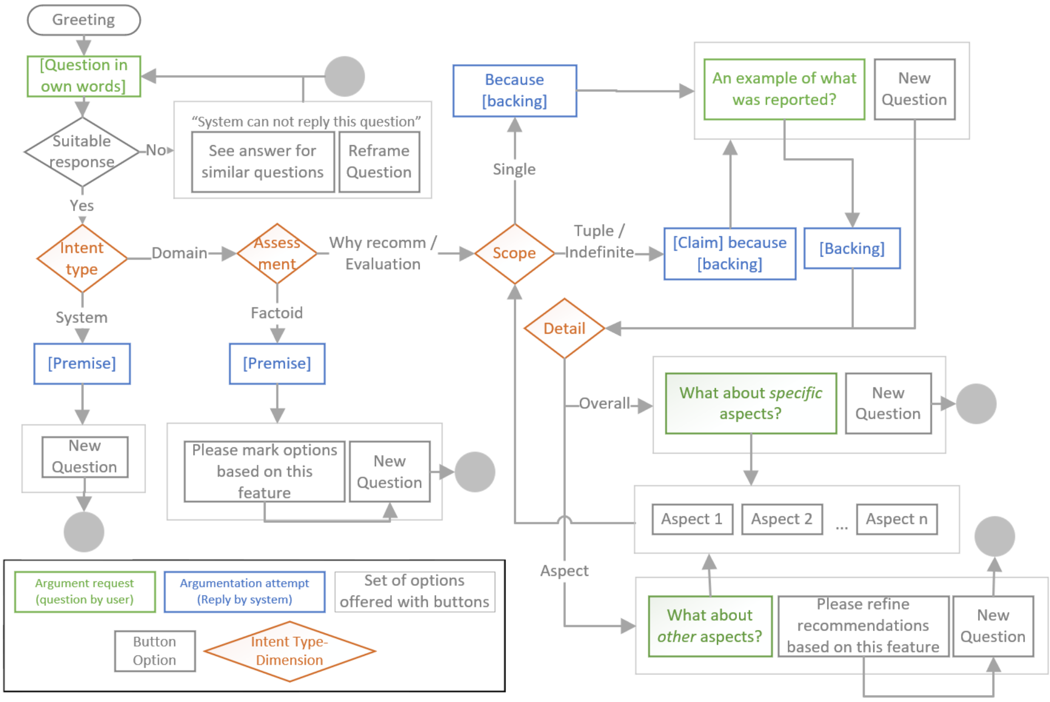

Greeting
[Question in own words]
Suitable response
No
Yes
"System can not reply this question"
See answer for similar questions
Reframe Question
Intent type
Domain
System
Assess ment
Why recomm / Evaluation
Factoid
Scope
Single
Tuple / Indefinite
Because [backing]
An example of what was reported?
New Question
[Claim] because [backing]
[Backing]
Detail
Overall
Aspect
[Premise]
[Premise]
New Question
Please mark options based on this feature
New Question
What about specific aspects?
New Question
Aspect 1
Aspect 2
...
Aspect n
What about other aspects?
Please refine recommendations based on this feature
New Question
Argument request (question by user)
Argumentation attempt (Reply by system)
Set of options offered with buttons
Button Option
Intent Type-Dimension

Figure 8: Proposed dialog management policy for conversational explanations in review-based RS

non-explicit items to be compared (indefinite) clearly outnumber those that do make them explicit (tuple). This involves an important implementation challenge, given that the methods proposed to answer this type of questions, and which are capable of processing several opinions on multiple items, are very scarce [44]. Furthermore, it should be noted that many of these questions also did not indicate specific features for evaluation (the fourth most frequent type of intention, see Table 2), so not only calculating the answer is challenging, but also how to communicate it briefly.

In this regard, we observed that while most of the questions were aspect-based utterances, an important portion also asked for overall assessment of the hotel(s). Here, an adequate balance must be maintained between relevance of the response (information about user's preferred features should be provided) and brevity. Guidelines from [46] recommend responding with concise utterances in the first place, and then enable the possibility to expand the information if needed, which could be facilitated by providing the option to choose specific aspects to dive into further details. System could also use this implicit indication of preferences to recalculate recommendations, as discussed for factoid questions.

Additionally, as expected, users not only generated standalone questions, but also follow-up questions, which confirms our expectation that an interactive QA approach would be appropriate to keep track of context and previously referred entities. Although creating a system able to respond to all possible questions is yet unrealistic [46], we suggest acknowledging "the system cannot answer this question" when the exact request cannot be answered. However, we suggest enabling the option of getting a response on a related feature, provided that the questioned aspect is within a reasonable range of similarity to those addressed by the system (e.g. "criminality rate" is related to the aspect "safety").

Finally and based on our observations, we adapted the scheme for interactive RS explanations proposed by [22], and extended it to support conversational argumentation, as well as system actions that could be triggered during the conversation. Figure 8 shows the proposed dialog management policy.

7 CONCLUSION AND FUTURE WORK

In this paper we have explored the design of a dialog management policy for explanations as conversational argumentation in review-based RS, conducted a WoOz study to assess the types of questions users might ask, and modeled user intents that could be used for the implementation of an explanatory CA in a hotel RS.

While the results obtained allowed us to gain a first insight into the type of questions that users would ask in the context under study, we acknowledge that a larger sample of participants would allow us to establish with more certainty the range of possible questions and reactions to the system's responses by users. Thus, we plan as future work, to continue with the implementation of the proposed methods for both automatic recognition of intents and generation of responses, as well as the implementation of the proposed policy within a dialog system, so that conversations can be collected on a larger scale. The above would also allow us to assess users' perception of the proposed solution, as compared, for example, to RS with static, or interactive but non-conversational explanations.

ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18*. 1–18.

[2] Abdallah Arioua and Madalina Croitoru. 2015. Formalizing Explanatory Dialogues. *Scalable Uncertainty Management* (2015), 282–297.

[3] Abdallah Arioua, Madalina Croitoru, Laura Papaleo, Nathalie Pernelle, and Swan Rocher. 2016. On the Explanation of SameAs Statements Using Argumentation. *Scalable Uncertainty Management* (2016), 51–66. https://doi.org/10.1007/978-3-319-45856-4_4

[4] Roland Bader, Wolfgang Woerndl, Andreas Karitnig, and Gerhard Leitner. 2012. Designing an explanation interface for proactive recommendations in automotive scenarios. In *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization (UMAP'11)*. 92–104.

[5] Sabrina Barko-Sherif, David Elsweiler, and Morgan Harvey. 2020. Conversational Agents for Recipe Recommendation. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 73–82. https://doi.org/10.1145/3343413.3377967

[6] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 717–725.

[7] Jamal Bentahar, Bernard Moulin, and Micheline Belanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review* 33, 3 (2010), 211–259.

[8] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP*. 5480–5494. https://doi.org/10.18653/v1/2020.emnlp-main.442

[9] Marco De Boni and Suresh Manandhar. 2020. Implementing clarification dialogues in open domain question answering. *Natural Language Engineering* 11, 4 (2020), 343–361. https://doi.org/10.1017/S1351324905003682

[10] Dimitrios Buhalis and Emily Siaw Yen Cheng. 2020. Exploring the Use of Chatbots in Hotels: Technology Providers Perspective. *Information and Communication Technologies in Tourism* (2020), 231–242. https://doi.org/10.1007/978-3-030-36737-4_19

[11] Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. In *Artif. Intell.*, Vol. 170. 925–952.

[12] Li Chen and Pearl Pu. 2014. Critiquing-based recommenders: survey and emerging trends. 22, 1–2 (2014), 3085–3094.

[13] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 16*. 815–824. https://doi.org/10.1145/2939672.2939746

[14] Oana Cocarascu, Antonio Rago, , and Francesca Toni. 2019. Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*.

[15] Nils Dahlback, Arne Jonsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st Int. Conference on Intelligent User Interface*. 193–200.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019).

[17] Rebecca Eynon, Chris Davies, and Wayne Holmes. 2012. Wizard of Oz for Multimodal Interfaces Design: Deployment Considerations. In *Proceedings of the 8th International Conference on Networked Learning*. 66–73.

[18] Miguel Grinberg. 2020. Socket.IO. (2020). https://github.com/miguelgrinberg/Flask-SocketIO

[19] Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. In *Computational Linguistics 43*, Vol. 1. 125–179.

[20] Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*.

[21] Diana C Hernandez-Bocanegra and Juergen Ziegler. 2020. Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics. *Journal of Interactive Media* 19, 3 (2020), 81–200. https://doi.org/10.1515/icom-2020-0021

[22] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Effects of interactivity and presentation on review-based explanations for recommendations. In *arXiv:2105.11794*. http://arxiv.org/abs/2105.11794

[23] Denis J. Hilton. 1990. Conversational processes and causal explanation. 107, 1 (1990), 65–81.

[24] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A Survey on Conversational Recommender Systems. 1–35. https://doi.org/abs/2004.00646

[25] Nitin Jindal and Bing Liu. 2006. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 06*. 244–251. https://doi.org/10.1145/1148170.1148215

[26] John F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Information Applications. In *Transactions on Office Information Systems*, Vol. 2. 26–41.

[27] Lisa Klein. 1998. Evaluating the Potential of InteractiveMedia through a New Lens: Search versus Experience Goods. In *Journal of Business Research*, Vol. 41. 195–203.

[28] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. In *User Modeling and User-Adapted Interaction*. 441–504.

[29] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI 19)*. ACM, 379–390.

[30] Béatrice Lamche, Ugur Adigüzel, and Wolfgang Wörndl. 2012. Interactive explanations in mobile shopping recommender systems. In *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE'14), held in conjunction with the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP'14)*. 92–104.

[31] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. Klagenfurt, Germany: SSOAR.

[32] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *32nd Conference on Neural Information Processing Systems, NeurIPS 2018*. 9725–9735.

[33] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 9042 (2020), 1–15. https://doi.org/10.1145/3313831.3376590

[34] Nathalie Rose Lim, Patrick Saint-Dizier, , and Rachel Roxas. 2009. Some challenges in the design of comparative and evaluative question answering systems. In *In Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions - KRAQ 09*. 15–18. https://doi.org/10.3115/1697288.1697292

[35] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement* 27 (1990), 247–266.

[36] Yuping Liu and L J Shrum. 2002. What Is Interactivity and Is It Always Such a Good Thing? Implications of Definition, Person, and Situation for the Influence of Interactivity on Advertising Effectiveness. *Journal of Advertising* 31, 4 (2002), 53–64.

[37] Benedikt Loepp, Katja Herrmanny, and Juergen Ziegler. 2015. Blended Recommending: Integrating Interactive Information Filtering and Algorithmic Recommender Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15*. 975–984.

[38] Benedikt Loepp, Tim Hussein, and Juergen Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI 14*. 3085–3094.

[39] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019*. 1–9.

[40] Bella Martin and Bruce Hanington. 2012. *Universal Methods of Design*. Rockport Publishers, Beverly, MA.

[41] Philipp Mayring. 2014. Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution. (2014). Klagenfurt, Germany: SSOAR.

[42] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. In *Information Systems Research*, Vol. 13.

[43] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* (2018).

[44] Amit Mishra and Sanjay Kumar Jain. 2015. An Approach for Sentiment analysis of Complex Comparative Opinion Why Type Questions Asked on Product Review Sites. *Computational Linguistics and Intelligent Text Processing Springer LNCS* 9042 (2015), 257–271.

[45] Christof Monz. 2003. Document Retrieval in the Context of Question Answering. In *Proceedings of the 25th European conference on IR research*. 571–579.

[46] Robert J. Moore and Raphael Arar. 2018. Conversational UX Design: An Introduction. *Studies in Conversational UX Design* (2018), 1–16. https://doi.org/10.1007/978-3-319-95579-7_1 Springer International Publishing.

[47] Emanuela Moreale and Maria Vargas-Vera. 2004. A Question-Answering System Using Argumentation. *MICAI 2004: Advances in Artificial Intelligence* (2004), 400–409. https://doi.org/10.1007/978-3-540-24694-7_41

[48] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Intelligent User Interfaces (IUI 16)*, Vol. 2. 256–260.

[49] Philip J. Nelson. 1981. Consumer Information and Advertising. In *Economics of Information*. 42–77.

[50] Hans Dybkjaer Niels Ole Bernsen and Laila Dybkjaer. 1993. Wizard of Oz prototyping: How and when. In *In CCI Working Papers in Cognitive Science and HCI*.

[51] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 135–143. https://doi.org/10.1145/3349537.3351899

[52] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems - RecSys 11*. 157–164.

[53] Silvia Quarteroni and Suresh Manandhar. 2008. Designing an interactive open-domain question answering system. *Natural Language Engineering* 15, 1 (2008), 73–95. https://doi.org/10.1017/S1351324908004919

[54] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, and Francesca Toni. 2020. Argumentation as a Framework for Interactive Explanations for Recommendations. In *Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning*. 805–815.

[55] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and Metrics for Cold-Start Recommendations. In *Proceedings of SIGIR 2002*. 253–260.

[56] Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. 34, 2 (2020), 235–250.

[57] Ronnie Taib and Natalie Ruiz. 2007. Wizard of Oz for Multimodal Interfaces Design: Deployment Considerations. In *Human-Computer Interaction. Interaction Design and Usability*. 232–241.

[58] Nava Tintarev. 2007. Explanations of recommendations. *Proceedings of the 2007 ACM conference on Recommender systems, RecSys 07* (2007), 203–206.

[59] Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *15th International Conference on Intelligent Text Processing and Computational Linguistics*. 115–127.

[60] Douglas Walton. 2000. The Place of Dialogue Theory in Logic, Computer Science and Communication Studies. 123 (2000), 327–346.

[61] Douglas Walton. 2004. A new dialectical theory of explanation. 7, 1 (2004), 71–89.

[62] Douglas Walton. 2011. A dialogue system specification for explanation. 182, 3 (2011), 349–374.

[63] Douglas Walton and Erik C. W. Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, New York.

[64] Nan Wang, Hongning Wang, Yiling Jia, , and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 18*. 165–174.

[65] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. 62, 6 (2019), 70–79.

[66] Markus Zanker and Martin Schoberegger. 2014. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*. 33–36.

[67] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 177–186. https://doi.org/10.1145/3269206.3271776

[68] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*. 83–92.

Paper 5

The following paper is reused from:

- Hernandez-Bocanegra, D.C. & Ziegler, J. (2021). ConvEx-DS: A dataset for conversational explanations in recommender systems. In *Proceedings of IntRS 21: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 1-18. url: http://ceur-ws.org/Vol-2948/paper1.pdf

ConvEx-DS: A dataset for conversational explanations in recommender systems

Diana C. Hernandez-Bocanegra, Jürgen Ziegler

University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany

Abstract

Conversational explanations are a novel and promising means to support users' understanding of the items proposed by a recommender system (RS). Providing details about items and the reasons why they are recommended in a conversational, language-based style allows users to question recommendations in a flexible, user-controlled manner, which may increase the perceived transparency of the system. However, little is known about the impact and implications of providing such explanations, using for example a conversational agent (CA). In particular, there is a lack of datasets that facilitate the implementation of dialog systems with explanatory purposes in RS. In this paper we validate the suitability of an intent model for explanations in the domain of hotels, collecting and annotating 1806 questions asked by study participants, and addressing the perceived helpfulness of the responses generated by an explainable RS using such intent model. Thus, we release an English dataset (ConvEx-DS), containing intent annotations of users' questions, which can be used to train intent classifiers, and to implement a dialog system with explanatory purpose in the domain of hotels.

Keywords

Recommender systems, explanations, conversational agent, user study, dataset

1. Introduction

Providing explanations of the rationale behind a recommendation can bring several benefits to recommender systems (RS), by increasing users' perception of transparency, effectiveness, and trust [1]. Although most explanations in RS are presented statically (i.e., using a fixed display in a single step), recent work has shown that providing interactive options for obtaining explanatory information can positively influence users' perception of RS [2]. Interactive options in explanations allow users to take control over the desired level of detail of the explanatory information, by means of a two-way communication, where users can indicate to the system the most relevant aspects on which explanations should focus. However, these possibilities are mostly limited to click-based options. Another kind of interactive approach to explanations is the conversational approach, in which users can express their questions in their own words. However, this has been, so far, much less explored.

While conversational approaches have already gained some attention in explainable artificial intelligence (XAI), and formal models of conversational explanations have been proposed to

IntRS'21: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 25, 2021, Virtual Event

diana.hernandez-bocanegra@uni-due.de (D. C. Hernandez-Bocanegra); juergen.ziegler@uni-due.de (J. Ziegler)
0000-0002-1773-2633 (D. C. Hernandez-Bocanegra); 0000-0001-9603-5272 (J. Ziegler)

this end [3, 4, 5], little is known about the type of explanation-related questions users would ask to a RS. Although several datasets exist that support the development of dialog and question-answering (QA) systems, these are generally focused on open domain-search (e.g. [6, 7]), or specific processes such as flight or hotel booking, without a focus on explanatory interaction as such. In particular, and to our knowledge, there are no publicly available datasets intended to support the development of an explanatory dialog system for RS, specifically, there are as yet no datasets for detecting the user's intent expressed in a question. We therefore collected 1806 questions that users asked to a RS and annotated them with intents according to an intent classification scheme developed for this purpose by [8]. The dataset contains questions about hotel recommendations and supports machine-learning based intent classification for explanatory conversational agents (CA).

Query intents are often characterized by means of intent classification schemes, which usually involve multiple dimensions (e.g. [9, 10]). This approach can facilitate the implementation of automatic intent detection procedures (i.e. those allowing to identify what information a user desires [10], so a proper answer can be generated), since detection can be solved by splitting the task into several less complex text classification tasks, one per each dimension. However, to implement text classifiers based on the intent model of [8], we still faced a challenge: even though some existing datasets could be useful for classifying values of the proposed dimensions (e.g. comparison or assessment), some relevant dimension values (or classes) are not annotated in those datasets, as discussed in depth in section 2.4.

We extend previous work of [8], who collected a small set of 82 human-generated questions about recommended items through a Wizard of Oz (WoOz) study. Their proposed model addresses two entities: *hotel* and *hotel feature*, and two main intent types: *system-related* intents (related to the algorithm, or the system input) and *domain-related* intents (related to hotels and their features). In turn, the *domain-related* type consists of the following dimensions, with several values each: *comparison* (a question could be comparative or not), *assessment* (whether question refers to facts, to a subjective evaluation or the reasons why an item is recommended), *detail* (whether the question refers to a single aspect or to the entire item), and *scope* (whether the question is about a single item, several items, or to the whole set of recommendations). The authors argue that an intent can be considered as a combination of values of these dimensions, and that reasonable answers can be generated when using such a scheme. For instance, the intent expressed by "Why are the rooms at Hotel X great?" would be: non-comparative (comparison) / why-recommended (assessment) / aspect (detail) / single (scope); and in consequence - assuming a review-based explanation method -, a possible answer could be: "because 96% of opinions about rooms are positive".

Given that the dimension-based intent model [8] was derived on a very small data set, it is still necessary to evaluate the validity of this proposal on a larger scale, i.e., the extent to which the model is able to accurately represent user intents given a larger set of questions. In particular, as an indirect measure of validity, we set out to evaluate perceived helpfulness of the responses generated by an RS implementing the intent model, under the assumption that if the system has adequately recognized the user's intent, it is able to generate a response that approximates the user's information need, and thus be considered, to some extent, helpful. Therefore, in this paper we aim to answer: **RQ1**: How valid is the dimension-based intent model proposed by [8], when taking into account a larger number of user-generated questions?

To this end, we collected a corpus of 1806 questions, and evaluated the perceived helpfulness by users of the answers generated by the system we implemented for this purpose. Additionally, we annotated the intent of the collected questions, using guidelines inspired by the intent model definition. Our aim was twofold: 1) to train classifiers with a view to future developments and further empirical validation of the conversational approach, and 2) to further validate whether the intent model could generalize to a larger scale. More specifically: **RQ2**: To what extent the collected questions could be consistently classified by human annotators?

To answer this question, we calculated inter-annotator agreement and assessed the pattern of questions where agreement was low, as well as particular observations that arose during the annotation process (detailed in section 4).

Finally, we consolidated the intent gold standard for each question, and validated the performance of intent detection procedures trained using the final annotated corpus. More specifically: **RQ3**: To what extent does the intent classification perform better when trained on our annotated dataset, compared to the auxiliary datasets we used during corpus collection?

To answer our research questions, we implemented an explanatory RS, which could interpret user queries and provide answers based on the underlying RS algorithm used ([11]), and text classifiers for the different dimensions, based on the state-of-the-art natural language processing (NLP) model BERT [12]. These classifiers were initially trained on *auxiliary* datasets that could be useful for detecting certain (but not all) dimension values (as detailed in section 3.1). We then conducted a user study aiming both to collect a large number of user queries, and to evaluate the perceived helpfulness by users of system generated answers. Details of system implementation and corpus collection procedure are addressed in section 3. Finally, the contributions of this paper can be summarized as follows:

- We release ConvEx-DS [1] (**Conv**ersational **Ex**planations **D**ata**S**et), consisting of 1806 user questions with explanatory purpose in the domain of hotels, with question intent annotations, which can facilitate the development of explanatory dialog systems in RS.
- We implemented a RS that generates answers to these questions, and tested the user-perceived helpfulness of system generated answers.

2. Related work

2.1. Explanations in RS

Providing explanations for recommended items can serve different purposes. Explanations may enable users to to understand the suitability of the recommendations, to understand why an item was recommended, or they may assist users in their decision making. Among the most popular approaches are the methods based on collaborative filtering (e.g. “Your neighbors’ ratings for this movie” [13]), as well as content-based methods that allow feature-based explanations, showing users how relevant item features match their preferences (e.g. [14]). On the other hand, review-based explanations usually show summaries of the positive and negative opinions about items (e.g. [15, 16, 17]). Our work is related to the latter approach, and our implemented

[1] ConvEx-DS can be downloaded at https://github.com/intsys-ude/Datasets/tree/main/ConvEx-DS

system uses the explanatory RS method proposed by [11], to generate both recommendations and explanations, based on ratings and customers' opinions.

2.2. Interactive and conversational explanations

In contrast to static approaches to explanations (which are dominant in RS and XAI overall [18]), interactive approaches seek to provide users with greater control over the explanatory components [2, 19], so that a better understanding of the reasons behind the recommendations can be achieved.

Moreover, conversational approaches to explanations take into account the social aspect of this process [20], where "someone explains something to someone" [21], through an exchange of questions and answers between the user and the system, as would occur in a human conversation. To this end, formal specifications and dialogue models of explanation (e.g. [3, 22, 5]) have been proposed as a theoretical basis for designing conversational explanations in intelligent systems. However, due to lack of sufficient empirical evaluation of such approaches [20, 4], it is still unclear how conversational explanatory interfaces should be conceived and designed in RS.

Recently, and inspired by dialog models of explanation [23], [8] proposed a dialog management policy and an user intent model, to implement a CA for explanatory purposes in a hotel RS. Our work builds on this model, and we extended this work by evaluating the intent model validity on a larger scale. While the prior work was based on the Wizard of Oz (WoOz) method for collecting user questions followed by explanations given by the experimenter, resulting in a set of 82 questions, in the present work we implemented a system to automatically generate answers, which allowed us to collect a larger number of questions (1806).

2.3. Intent detection and slot filling

We developed an RS system able to reply automatically to users' questions as part of an explanatory conversation. To this end, we set our focus on the natural language processing (NLP) tasks: intent detection and slot filling, key tasks for the development of dialog systems. Intent detection seeks to interpret the user' information need expressed through a query, while slot filling aims to detect which entities - and also features of an entity - the query refers to. The idea behind the intent concept is that user utterances within a dialogue can be framed within a finite and more limited set of possible dialogue acts. The most common approach for intent representation, in the open-search domain, is intent classification [10], that is, a query can be categorized according to a classification scheme, consisting of dimensions or categories, and their possible values, as in [9, 10]. The intent model by [8], on which we base our work, falls within this type of representation.

A large body of previous work has addressed the task of intent detection, both for open search domains (see [24, 10]) and task-oriented dialog systems, for processes such as flight booking, music search or e-banking, e.g. [25, 26, 27]. Methods proposed to solve these tasks range from conventional text classification methods, to more complex neural approaches, based on recurrent neural networks, attention-based mechanisms and transfer learning, to solve the intent detection and slot-filling tasks, both jointly and independently, and to extend the solutions to new domains. Since an in-depth comparison of the different approaches is beyond the scope of this paper, we refer readers to the survey on this matter by [28].

In particular, our work is related to the text classification approach, which leverages the representation of possible intents according to a classification scheme. According to this approach, the difficult task of intent detection can be divided into smaller text classification tasks, to detect the class that best represents a sentence according to each dimension. For this purpose, we implemented text classifiers using the state of art natural language processing model BERT [12]. As for the slot-filling task, and in line with [29], we solve it as a named entity recognition (NER) task. In our case, the entities to be recognized correspond to the names of the hotels about which the questions are asked. For this purpose, we use the NLTK toolkit [30].

2.4. Datasets for Intent detection

Benchmarking of intent-detection tasks is usually based on prominent datasets like ATIS [25] (Airline Travel Information System, containing queries related to flight searching), the MIT corpus [31] (queries to find movie information, or booking a restaurant), or the SNIPS dataset [26] (to develop digital assistants, involving tasks as asking for weather, or playing songs). To our knowledge, no public dataset has been published to support the development of dialog systems with explanatory purpose in RS. However, we investigated existing data sets that could contribute to classifying values along the different dimensions of the intent model.

Dimension comparison: Work by Jindal and Liu [32] addressed the identification of comparative sentences as a classification problem. The authors released a dataset with comparative and non-comparative sentences extracted from user reviews on electronic products, from forums involving comparison between brands or products, and from news articles on random topics. On the other hand, Panchenko et al. [33] released a dataset for comparative argument mining, involving 3 classes (better, worse or none), in domains like computer science, food or electronics. This dataset allows automatic detection of comparative sentences where entities to compare are explicitly mentioned (e.g. “Python is better suited for data analysis than MATLAB”), while superlative sentences like “which is the best option?” are not considered as comparative. The above was problematic for our purposes, since most comparative questions in the WoOz [8] set are precisely superlative. Consequently, we opted to use Jindal and Liu (see section 3.1).

Dimension assessment: Bjerva et al. [34] released SubjQA, a dataset for several domains (including hotels), which can be used to detect subjectivity of questions, in QA tasks. This dataset includes annotations of subjective and non-subjective classes, which can be leveraged to classify *evaluation* and *factoid* questions, according to the intent model by [8]). This dataset does not involve questions of the type *why-recommended*.

Dimension detail: While most aspect-based approaches involve the detection of an aspect or specific feature addressed in a sentence (e.g., room, facilities), as in [35, 2], detecting the absence of aspect is not usually addressed. Consequently, to our knowledge, no dataset involves annotation of sentences that addressed the quality of an overall item (e.g., "how good is Hotel x?"), in contrast to aspect-based sentences. Therefore, we used sentences collected in the WoOz study and the classification “detail”, as described in Section 3.1.

Dimension scope: To our knowledge, there is no dataset for the detection of the scope dimension. However, the values under this dimension can be inferred from entity detection (particularly hotels), for which NER can be used.

3. Corpus collection

Aiming to validate the intent model proposed by [8], we implemented and tested a conversational RS, consisting of a natural language understanding (NLU) module, which interprets questions with explanatory purpose written by users, and a module to generate answers consistent with the review-based recommendation method on which the RS is based. The development of our system and the corresponding user study involved a process consisting of several iterations. After every iteration, participants were asked to interact with the latest version of our system, so results of each iteration were used to improve the system to be tested in the next iteration. This was done in order to improve the performance of the classifiers, and to include new methods of response generation, for example to respond to intents that were not initially implemented, given their low frequency among all the questions asked by users. In addition to collecting participants' questions, we also captured their perception of the helpfulness of the answers generated by the system. Details of the methods implemented and the user study below.

3.1. Intent detection: methods and datasets

We divided the intent detection task into a set of three classification tasks (one for each of the dimensions: *comparison*, *assessment* and *detail*), and one NER task (for the detection of "hotel" entities, which allowed us to infer the *scope* dimension). Thus, the final detected intent corresponds to the combination of the values detected of each dimension. Thus, for example, the intent of the sentence "how good is the service at Hotel X" should be detected as: non-comparative (comparison) / evaluation (assessment) / aspect (detail) / single (scope).

BERT-based Text classifiers: We trained BERT classifiers [12], one for each dimension (comparison, assessment and detail), using a 12-layer model (*BertForSequenceClassification*, bert-base-uncased), batch size 32, and Adam optimizer (learning rate = 2e-5, epsilon = 1e-8). Classifiers for comparison and assessment converged after 4 epochs, while for detail 5 epochs were needed. Datasets were split randomly into training (80%) and test (20%) during the training phase. In order to avoid overfitting, the most represented class was downsampled (randomly) to approximate the size of the less represented class, which was slightly upsampled (randomly) to fit round numbers like 1000 and 500. In the case of the detail dimension, due to the small size of the auxiliary dataset, both classes were increased to 100 instances each (described below). Datasets (original and balanced) sizes are reported in table 1.

Dimension comparison: To train the classifier, we used the dataset by Jindal and Liu [32], which involves 5 classes (non-equal gradable, equative, superlative, non-gradable and non-comparative), all except the last one correspond to a detailed level of granularity for the sentences considered as comparative, which we believe is not necessary for our purposes. Thus, we grouped the sentences of the comparative classes (non-equal gradable, equative, superlative, and non-gradable), under a single *comparative* class.

Dimension assessment: We used the dataset by Bjerva et al. [34], specifically the one corresponding to the domain of hotels. Dataset includes an annotation whether the sentence is subjective or not, which we used to classify questions as evaluative and factoid, respectively. As this dataset does not involve the class *why-recommended*, we included a handcrafted validation, so subjective questions including the word "why" were regarded as such.

Table 1
Size and distribution of datasets used to train initial classifiers, implementation used during corpus collection phase.

	Dataset size	
Comparison [32]	**Original**	**Balanced**
Comparative	853	1000
Non-Comparative	7200	1000
Subjectivity [34]	**Original**	**Balanced**
Subjective	2706	500
Non-subjective	488	500
Detail [8]	**Original**	**Augmented, balanced**
Aspect	58	100
Overall	22	100

Dimension detail: Here, we leveraged questions collected by [8] in their WoOz study. As the size was extremely low, we used an augmentation technique, to generate synthetically new sentences from those in the WoOz dataset, altering some words, such as hotel names or aspects. Additionally, after initial iterations of the user study, we manually classified the collected questions written by participants as *aspect* or *overall*, added new sentences from the less represented class (i.e. the *overall*) to the dataset and retrained the *detail* classifier, so the risk of overfitting due to imbalanced classes and augmentation techniques could be decreased in the next iteration.

Dimension scope and entity 'hotels': First, we identify the entities (hotels) mentioned in the sentence. For this NER task, we used procedures from the library NLTK [30] to identify the entities (particularly the tokenizer and the part of the sentence (POS) tag methods). Then, we inferred the scope value depending on the number of entities recognized: *single* for one entity, *tuple* for more than two, and *indefinite* if no entity was found. An special case are the anaphoras regarding the entity [36], e.g. the sentence "how is the service at *this* hotel?" might refer to a previous hotel mentioned in the previous question or its answer. Usually, these situations are handled by the dialog system, which is in charge of keeping track of context. As a solution, when no entity was detected, but the sentence included a determiner such as 'this', 'these', 'those', 'its', 'their', etc, and if an entity was recognized or included in the previous question or its answer, the sentence was marked as *single* or *tuple*.

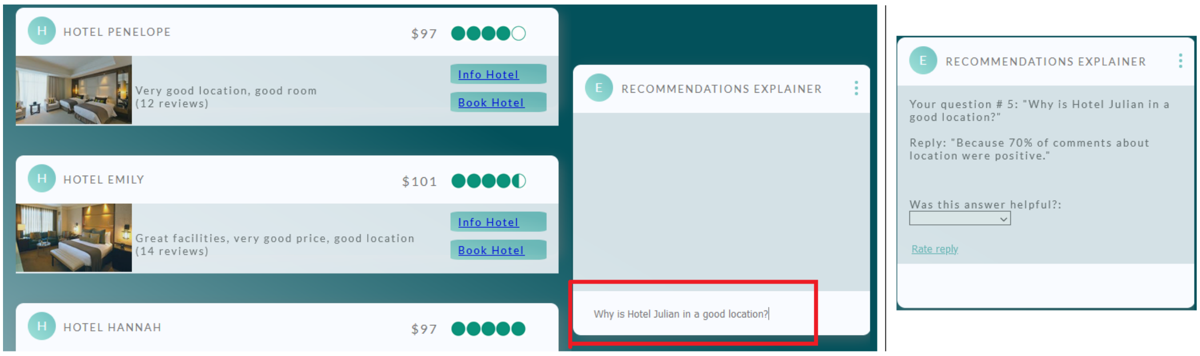
Entity 'hotel feature': We used the aspect-based detection methods implemented by [2], which use BERT classifiers and the ArguAna dataset [35], to detect aspect and hotel features addressed in user questions.

3.2. Explainable RS

We used the review-based RS developed by [2], which implements the matrix factorization model proposed by [11], in combination with sentiment-based aspect detection methods, using the state of art NLP model BERT [12], in order to provide aspect based arguments. We also use the personalization mechanism described by [2], which uses the aspects reported as preferred by participants in the study survey, to generate personalized recommendations.

Answer generation module: We implemented a module to generate replies based on the intent detected, and based on the type of argumentative responses proposed by [2]. According to this proposal, *factoid* questions could be replied with Y/N or a value (e.g. check in times) based on metadata. As for *evaluation* or *why-recommended* questions, replies were based on the aggregation of positive or negative opinions regarding an aspect (if question was *aspect* based), or the most important aspects for the participant (in case question was *overall*). This aggregation of opinions was calculated based on the hotels the question was about. If the question was comparative, the system calculated which hotel was better among a tuple, or the best in general (scope *indefinite*), based on the aggregation of the opinions. These are some examples of the type of responses generated by the system: Q: "Why does Hotel Hannah have the highest rating?", A: "Because of the positive comments reported regarding the aspects that matter most to you: 86% about location, and 85% about price."; Q: "Which hotel is best, Hotel Lily, Hotel Amelia or Hotel Hannah?", A: "Hotel Lily has better comments on the aspects that are most important to you (location, facilities, staff). However, Hotel Amelia has better comments about room, price."; Q: "Hotel Amelia is described as having a great room, what makes it great?, A: "Comments about rooms are mostly positive (90%).".

3.3. User study



HOTEL PENELOPE $97 ●●●●○

Very good location, good room
(12 reviews)

HOTEL EMILY $101 ●●●●◐

Great facilities, very good price, good location
(14 reviews)

HOTEL HANNAH $97 ●●●●●

Why is Hotel Julian in a good location?

Your question # 5: "Why is Hotel Julian in a good location?"

Reply: "Because 70% of comments about location were positive."

Was this answer helpful?:

Figure 1: Interface system used for corpus collection. Left, list of recommendations and box to write questions (highlighted in red). Right, system shows answer and requests users to rate helpfulness

Participants: We recruited 298 participants (209 female, mean age 30.42 and range between 18 and 63) through the crowdsourcing platform Prolific. We restricted the task to workers in the U.S and the U.K., with an approval rate greater than 98%. Participants were rewarded with 1.5 plus a bonus up to 0.30 depending on the quality of their response to the question "Why did you choose this hotel?" set at the end of the survey, aiming to achieve a more motivated hotel choice by participants, and encouraging effective interaction with the system. Time devoted to the task (in minutes): M=7.31, SD= 4.97. Questions asked per participant: M=5.99, SD=2.58.

We applied a quality check to select participants with quality survey responses (we included attention checks in the survey, e.g. "This is an attention check. Please click here the option 'yes'"). Users were told in the instructions that at least 5 questions were required as a prerequisite for payment, as well as correct answers for the attention checks (2). We discarded participants with at least 1 failed attention check, or no effective interaction with the system, i.e. if users did not ask questions to the system. Thus, the responses of 41 of the 339 initial participants were discarded and not paid (final sample: 298 subjects).

Procedure: Users were asked to report a list of their 5 most important aspects when looking for a hotel, sorted by importance. Then we presented participants with instructions: 1. They would be presented with a list of 10 hotels with the results of a hypothetical search for hotels already performed using a RS (i.e., no filters were offered to search for hotels). 2. They could consult general hotel information (photos, reviews, etc., by clicking on "Info Hotel"), but indicated that we were more interested in knowing their questions about the reasons why these hotels were recommended, stating that "The aim of the system is to provide explanations based on your questions" (we aimed here to prevent the user from asking questions about other processes, such as booking assistance). 3. They should write each question (in their own words) at the bottom of the explanation box (highlighted in red in the example), and click enter to send (see Fig. 1 left). 4. Next, the system would present the answer to their question, and a drop-down list to evaluate how helpful you think the answer was (with values from "Strongly disagree" to "Strongly Agree"). They had to choose a value from the list and click on the "Rate Reply" link, continue with your next question, and repeat until they complete at least 5 questions (Fig. 1 right). 6. Once they finished, they had to indicate which hotel they would finally choose by clicking the button "book hotel". 7. Back on the survey, they had to describe why they chose that hotel, we stated that a bonus would be paid depending on the quality of this response. A reminder of these instructions was included in the app, so it would be easier for users to remember them. After instructions and before the task, we presented a cover story, to establish a common starting point in terms of travel motivation (a holiday trip). The question used to rate the usefulness of the system's answers was: *"Was this answer helpful?"*, and reply was measured with a 1-5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

4. Corpus annotation

4.1. Intent type annotation

First, sentences where classified according to the classes: domain-related intents (regarding hotels and their features), and system-related intents (regarding the algorithm, the system Input, or system functionalities). [8] reported that domain questions clearly outnumbered system questions, so the research team members annotated this class instead of crowdsourcing workers (98.3% agreement), as the low number of system questions could lead to the category being ignored in the crowdsourcing setup. Disagreements were resolved in joint meetings.

4.2. Dimension-based annotation

Only domain-related sentences were used for the dimension-based annotation. We collected annotations for comparison, assessment and detail as independent tasks. The dimension scope was not annotated under the proposed procedure (is not a classification task but a NER task).

Annotators and crowdsourcing setup: Every sentence was annotated by 3 annotators: one belonging to the research team, and the other two crowdsourced on the Prolific platform. We divided the set of questions into 19 blocks of 100 sentences each, and every block had to be annotated for each dimension separately, to mitigate the fatigue associated with a longer list of questions, which could affect the participant's performance. Each block included 4 attention checks (e.g. "This is an attention check. Please click here the option 'comparative'"). Participants were warned that failing this check or not completing the list of 100 questions would lead to rejection and non-payment of the task. We also included questions from the examples provided in the guidelines within the blocks, for a subsequent attention check (failing this check led to rejection of the block for the agreement and final gold standard).

The research team annotator annotated all blocks for the three dimensions, while different crowdsourcing workers could annotate different blocks for different dimensions. Same annotator did not annotate the same block for the same dimension more than once. This way we ensured that each sentence was annotated by 3 different people, for each dimension.

Procedure: Once participants took the task in Prolific, they had to read the instructions of the task (annotation guidelines), and then open the annotation application (where annotation guidelines remained visible). Once the end (100 questions) was reached, the user was prompted to return to the main survey, and to report observations or difficulties.

Annotation application: We developed a simple annotation application, in which annotators could select the class to which a question belongs, according to each dimension. The user interface consisted of a single page, showing: at the top, a reminder of the guidelines for annotation; at the bottom, the consecutive number of the question (so that the user could note its progress, e.g. 2 out of 100), the question, a checkbox to indicate the class to which the sentence belonged, and a "Next question" button.

Participants, and selection of valid submissions: 92 participants performed the annotation task using the platform Prolific. We restricted the task to workers in the U.S and the U.K., approval rate greater than 98%. Participants who annotated comparison blocks were rewarded with 1.25, assessment blocks with 1.50, and detail blocks with 1.25. Differences in payment were due to the different devoted times in minutes for each dimension (comparison: M=9.85, SD=3.53, assessment: M=13.24, SD=5.86, detail: M=9.40, SD=2.69). Participants who failed the attention checks, or those who did not complete the task, were rejected and not paid (19 participants in total). None of the questions submitted by these participants were used in the final calculation of the gold standard, nor for the agreement score. As part of a subsequent quality check, we discarded participants and their submitted answers, if they failed to correctly classify questions that also appeared as examples in the instructions, although their submissions were paid (16 participants). No further criteria were used to discard blocks of user responses, as we were not to establish correct or incorrect answers, but to establish whether the elaborated guides were understood in a similar way by different users, and whether the classes established by the intent model fit the questions in the corpus. A final set of annotations by 57 Prolific workers was used for the calculation of Inter-rater reliability, and the deduction of gold standard.

Classifiers trained on ConvEx-DS: Bert model, batch size, Adam optimizer parameters, and splitting as reported in section 3.1. To avoid overfitting, the most represented class was downsampled (randomly), to approximate the size of the less represented class. Classifiers of comparison and assessment dimensions converged after 4 epochs, of dimension detail after 5.

5. Results

5.1. Helpfulness of system answers

Taking into account iterations 4 to 6 (most refined versions of the system) the system was able to generate an answer in 80.58% of the cases, and to partially recognize the intent or entities in 7.34% of the cases (thus asking the user to rephrase or indicate further information). Among the main reasons why the questions were not replied we found: complexity of the question or not information available to reply (31%), text that could be improved when replying factoid questions (23%), wrong intent classification (11%), and system errors (11%).

Figure 2 (middle) shows the perception of answers' helpfulness, according to ratings granted by study participants, across all iterations (M=3.58, SD=1.34). When taking into account only the last two iterations (which account for 63.85% of sentences collected, and involve the most refined versions of the system), we observed a greater perceived helpfulness (to M=3.70, SD=1.30). We considered as "non-helpful" responses those that were marked with the values "Strongly Disagree" and "Disagree" when participants were asked "Was the answer helpful?". We analyzed the responses given to those questions by the system and found that in 34% of cases, replies provided actually make sense, i.e. seemed a reasonable answer to the asked question. Among the reasons that caused the responses to be rated as non-helpful, we found: 30% due to misclassified intents or entities, 14% to system errors, 9% text that could be improved when replying factoid questions, 5% due to complexity of the question or not information to reply it, and 5% to specific aspects not addressed by the solution.

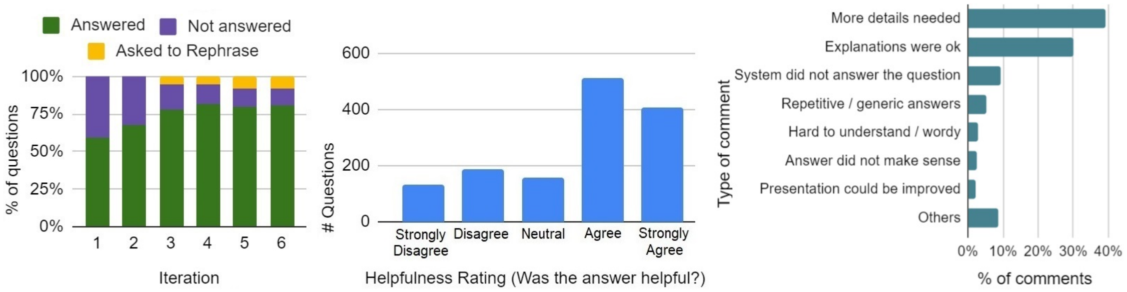


Figure 2: Left: Distribution of replied questions, across iterations. Middle: Histogram of helpfulness rating granted by users to answers generated by the system (all iterations). Right: Types of comments by participants during corpus collection, in regard to system answers.

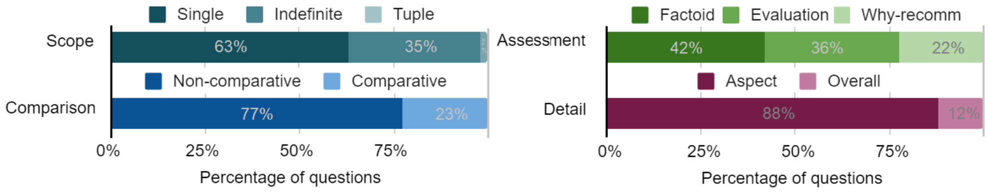


Figure 3: Distribution of questions in ConvEx-DS (domain-related intents).

Table 2

Inter-rater reliability of ConvEx-DS. *Fleiss' kappa* refers to each dimension, and the *% of full agreement* to each class, i.e. percentage of questions in which all annotators agreed on assigning that class).

Dimension	Fleiss' kappa	Class	% of full agreement
Comparison	0.72	Comparative	77.28%
		Non-comparative	86.86%
Assessment	0.65	Factoid	73.99%
		Evaluation	58.56%
		Why-recommended	66.42%
Detail	0.75	Aspect	95.73%
		Overall	66.82%

5.2. Annotation statistics

A total of 1836 questions were collected during the corpus collection step. 30 of those questions were discarded (nonsense statements, or highly ungrammatical), for a final set of 1806 of annotated questions. Of these, only 24 were annotated as system-related questions. Length of questions: characters M=39.2, SD=15.67, words M=7.35, SD=2.86. We found a Fleiss' kappa of 0.72 for *comparison*, of 0.65 for *assessment* and of 0.75 for *detail*, indicating a "substantial agreement" [37], for all three dimensions. As for classes with lower percentages of questions with full agreement, we identified the following main causes: **Dimension assessment:** - *Why-recommended* questions rated as subjective, given that adjectives like 'good' or 'great' are included in sentences, e.g. "why is hotel hannah location great?". - Questions that should be replied with a fact, but include adjectives that indicate subjectivity, e.g. "does hotel emily have any bad reviews?", "are there good transport links?", "which hotel best fits my needs?". - Questions with adjectives as 'cheap', 'expensive', 'close', 'near', 'far', which can be answered with either subjective or factoid responses, e.g. "Which is the cheapest hotel?", "is there an airport near any of these hotels?". - Questions of the type "what is ... like", e.g. "What is the room quality like at Hotel Emily?" (this type of questions were actually not addressed in instructions).

Dimension detail: - Concepts that were regarded as hotel aspects, e.g. value (Which is best value for money), ratings (What is the highest rating for Hotel Levi), reviews (Which hotel has the most reviews?), stars (Which hotels are 5 stars?).

Finally, we have detected some questions that could hardly fit in the planned classes, e.g. "How do you define expensive? Do you compare against facilities and what is included in the price?", "The Evelyn has 17 reviews and a positive feedback but scores lower than others with less reviews. Why is this?". However, we found this number to be rather low (16 questions).

5.3. Intent detection performance

Dimensions comparison, assessment and detail: To verify the performance of classifiers, we have calculated F1, a measure of classification accuracy. We tested accuracy in 3 different steps: 1) performance of models trained on auxiliary datasets [32, 34, 8], used for the system used in corpus collection. 2) We tested these models using our newly obtained annotated data, ConvEx-DS. 3) We trained and tested new classifiers, based entirely on ConvEx-DS. We report F1 scores for each dimension (*comparison*, *assessment* and *detail*). We reported weighted average,

Table 3
F1-scores (weighted average) of classifiers of different dimensions, trained and tested on both auxiliary datasets and ConvEx-DS.

Dimension	Dataset	F1
Comparison	Jindal and Liu [32] [Training, Testing]	0.87
	Jindal and Liu [32] [Training], ConvEx-DS [Testing]	0.88
	ConvEx-DS [Training, Testing]	0.92
Assessment	Bjerva et al. [34] [Training, Testing]	0.93
	Bjerva et al. [34] [Training], ConvEx-DS (without why-recomm) [Testing]	0.60
	ConvEx-DS [Training, Testing]	0.91
Detail	WoOz augmented [Training, Testing]	0.98
	WoOz augmented [Training], ConvEx-DS [Testing]	0.90
	ConvEx-DS [Training, Testing]	0.92

to take into account the contribution of each class, which in (2) is particularly unbalanced (no downsampling of the test set was done, since balanced data was pertinent only for training).

We detected that the classifier trained on Bjerva et al. [34], performed particularly poorly when tested with our annotated data (ConvEx-DS). Here, we detected that 32% of questions under "evaluation" class in ConvEx-DS but classified as "non-subjective" correspond to questions regarding indefinite or more than two hotels (e.g. "which hotel has the best facilities?"), 18% corresponded to adjectives like "close", "far", "expensive", and 14% to questions of the form "what is the food like?". As of factoid questions in ConvEx-DS classified as subjective, we found 33% of questions involving indefinite or more than two hotels, and 32% regarded questions of the form "does the hotel have...". In section 6 we discussed these findings in depth.

Dimension scope: Entities (hotels) addressed in sentences were detected using the NLTK library. In order to check the accuracy of the method, 2 members of the research team have checked the inferred entity for the collected corpus, and found that in 5.38% of cases, the inferences were wrong. Most of these cases corresponded to cities, or facilities recognized as entities, a drawback detected in early stages of corpus collection, thus additional validations were added to the procedure, so that these cases would not occur in future iterations.

6. Discussion

To date, creating dialog systems able to answer all possible users' questions remains unrealistic [38]. Nevertheless, we found that in the later iterations of our user study, our implemented system was able to answer a wide number of questions, or to ask users to rephrase or better specify their explanation need. However, since the ability to answer the questions is not a sufficient condition for concluding that a model of intent is valid, we set out to validate how helpful the answers were perceived by users, as an indirect measure of model validity, assuming that a correct intent detection would lead - to a certain extent - to the responses being perceived as helpful. In this respect, we found that system answers were perceived as predominantly helpful, thus answering **RQ1** positively. On the other hand, ratings of non-helpfulness did not necessarily imply that the queries did not match the detected intent. In fact, we found that almost one-third of responses rated as non-helpful fitted the question (i.e. made sense). After a review of participants' feedback on the system answers, we found that, although many users found them helpful or "ok," the main criticism was that some of the answers lacked sufficient detail. For instance, it was not enough to simply answer yes / no to factoid questions, but further details about the inquired feature were expected. As for the *evaluation* or *why-recommended*, participants reported that the percentages of positive and negative opinions were fine, but some also demanded examples of such opinions. The above is consistent with findings reported by [2], who found that perception of explanation sufficiency was greater when options were offered to obtain excerpts from customer reviews, and that the need for more detailed explanations may depend on individual characteristics, for instance, decision-making style: users with a predominant rational decision making style have a tendency to thoroughly explore information when making decisions [39].

In consequence, although the intent model seems appropriate to generate an initial or first level response, a dialog system implementation must go beyond this initial response, offering options to drill down into the details. Similarly, criticized aspects, such as repetitive or too generic answers, could also be mitigated with such a solution, since providing excerpts of customer reviews as answers would allow a balance between system-generated and customer-generated statements. An alternative in this respect is to provide natural language explanations based on customer reviews, using abstractive summarization techniques as in [40]. However, as reported [41], users seem to prefer explanations that include numerical anchors (e.g. percentages) in comparison to only based text summaries, since percentages may convey more compelling information, while summaries may be perceived as too imprecise.

In line with [8], we also found that questions related with *system-related* intents were clearly outnumbered by *domain-related* intents, showing that in the explanatory context of hotels RS, users usually do not formulate questions explicitly addressing the system or its algorithm. We believe that users are highly less interested in such details, due to the nature of the domain addressed. Hotels are experience goods (those which cannot be fully known until purchase [42]), for which an evaluation process is characterized by a greater reliance on word-of-mouth [42, 43], which may lead users to grant much more attention to item features and customers' opinions about it, rather than on the details of the algorithm or how their own profile is inferred.

In regard to the annotation task, we found a substantial agreement in all the annotated dimensions, as well as a very encouraging accuracy measure, when classifiers were trained on the ConvEx-DS, which leads us to conclude, that under the intent model and annotation guidelines based on [8], the questions could be, to a substantial extent, unequivocally classified, thus replying to our **RQ2**. We note, however, the challenge of addressing the dimension assessment. In this regard, we found that the main difficulty was to classify correctly questions that could be regarded as evaluation, given their subjective nature (including expressions like "how close/far"), but for which a factual-based response could be given (e.g. "100 meters from downtown"), a similar concern raised by [34]: "a subjective question may or may not be associated with a subjective answer". Additionally, questions like "why is hotel X good?" were often classified as *evaluation*, given their subjective nature (adjective "good" as an indicator of subjectivity), so they were regarded as similar to their evaluation counterpart ("how good is hotel X?"). However, we believe that the distinction "why good" should be kept separate from "how good", since in the former, the user challenges arguments already provided by the system (a recommendation, or its explanations), while in the later this is not necessarily the case.

As for **RQ3**, we found that intent classifiers perform better when trained on ConvEx-DS, compared to classifiers trained on the auxiliary datasets, but tested on ConvEx-DS. Here, the most striking case concerns the dataset for the detection of subjective questions (SubjQA) by Bjerva et al. [34]. The above in no way suggests anything problematic in the SubjQA itself, only that in comparison to ConvEx-DS (dimension "evaluation"), the two datasets measure rather different concepts. SubjQA addresses the subjectivity of the question asked, not whether the question involves an *evaluation* that might be subjective, as in ConvEx-DS. Thus, for example, "how is the food?" is classified as non-subjective under SubjQA, since it does not contain expressions indicating subjectivity. Thus, non-subjective under SubjQA does not necessarily imply factoid. In addition, classifiers trained in SubjQA do not work well with questions that involve some sort of comparison between multiple items, since the SubjQA only involves questions addressing single items, for which an answer could be found in a single review.

Limitations: Despite our motivating results, it is important to note the limitations imposed by the discussed approach. Addressing intent detection as a text classification problem, by means of an intent classification model, allows to provide answers that approximate the information need expressed by the user. However, the approach is insufficient when dealing with questions that are too specific, particularly in regard to factoid questions. Consequently, the development of a DS with explanatory purposes in RS should not only rely on the underlying RS algorithm, customer reviews or hotels metadata (as in our developed system), but should also integrate further sources of information, e.g. external location services, in order to provide very specific details, like surroundings, distances to places of interest or transport means, in case these are not found in customer reviews or metadata.

Also, our study setup for corpus collection was based on a question/answer sequence (with helpfulness rating in between), thus not necessarily resembling a fluid chatbot-style dialog, in which users might write utterances, such as greetings or thanks, expressions that could not be classified under the intent model. Therefore, we suggest the use of alternative mechanisms for the detection and treatment of such expressions.

7. Conclusions and future work

Based on our results, we conclude that the dimension-based intention model proposed by [8] is a valid approach to represent user queries in the context of explanatory RS. We also believe that ConvEx-DS can significantly contribute to the development of dialog systems that support conversational explanations in RS.

As future work, we plan to explore the users' perception of a RS, where further details and excerpts from customers reviews are provided during the explanatory conversation, aiming to increase the perceived helpfulness by users of the responses that our system is able to generate. Additionally, although questions in ConvEx-DS involve only one domain, we believe it can also be leveraged for the development of explanatory approaches in RS for other domains, especially those involving review-based recommendations. In this sense, we plan to explore recent NLP developments, particularly on transfer learning techniques, to obtain linguistic representations that can serve as a basis for similar domains, particularly those where customer reviews are also exploited, such as restaurants, movies and shopping.

Acknowledgments

This work was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group “User-Centred Social Media”.

References

[1] N. Tintarev, J. Masthoff, Explaining recommendations: Design and evaluation, in: Recommender Systems Handbook, Springer US, Boston, MA, 2015, p. 353–382.

[2] D. C. Hernandez-Bocanegra, J. Ziegler, Effects of interactivity and presentation on review-based explanations for recommendations, in: Human-Computer Interaction – INTERACT 2021, Springer International Publishing, 2021, pp. 597–618.

[3] D. Walton, The place of dialogue theory in logic, computer science and communication studies 123 (2000) 327–346.

[4] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019, 2019, p. 1–9.

[5] A. Rago, O. Cocarascu, C. Bechlivanidis, F. Toni, Argumentation as a framework for interactive explanations for recommendations, in: Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning, 2020, p. 805–815.

[6] E. Merdivan, D. Singh, S. Hanke, J. Kropf, A. Holzinger, M. Geist, Human annotated dialogues dataset for natural conversational agents, Appl. Sci 10 (2020) 1–16.

[7] A. Ritter, C. Cherry, W. B. Dolan, Data-driven response generation in social media, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, p. 583–593.

[8] D. C. Hernandez-Bocanegra, J. Ziegler, Conversational review-based explanations for recommender systems: Exploring users’ query behavior (in press), in: 3rd Conference on Conversational User Interfaces (CUI ’21), 2021.

[9] A. Broder, A taxonomy of web search, ACM SIGIR Forum 36 (2002) 3–10.

[10] S. Verberne, M. van der Heijden, M. Hinne, M. Sappelli, S. Koldijk, E. Hoenkamp, W. Kraaij, Reliability and validity of query intent assessments: Reliability and validity of query intent assessments, Journal of the American Society for Information Science and Technology 64 (2013) 2224–2237.

[11] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma., Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval, 2014, p. 83–92.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019).

[13] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, ACM, 2000, p. 241–250.

[14] J. Vig, S. Sen, J. Riedl, Tagsplanations: explaining recommendations using tags, in:

Proceedings of the 14th international conference on Intelligent User Interfaces, ACM, 2009, p. 47–56.

[15] K. I. Muhammad, A. Lawlor, B. Smyth, A live-user study of opinionated explanations for recommender systems, in: Intelligent User Interfaces (IUI 16), volume 2, 2016, p. 256–260.

[16] N. Wang, H. Wang, Y. Jia, , Y. Yin, Explainable recommendation via multi-task learning in opinionated text data, in: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 18, 2018, p. 165–174.

[17] D. C. Hernandez-Bocanegra, J. Ziegler, Explaining review-based recommendations: Effects of profile transparency, presentation style and user characteristics, Journal of Interactive Media 19 (2020) 181–200.

[18] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18, 2018, p. 1–18.

[19] K. Sokol, P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency 34 (2020) 235–250.

[20] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence (2018).

[21] D. J. Hilton, Conversational processes and causal explanation 107 (1990) 65–81.

[22] A. Arioua, M. Croitoru, Formalizing explanatory dialogues, Scalable Uncertainty Management (2015) 282–297.

[23] D. Walton, A dialogue system specification for explanation 182 (2011) 349–374.

[24] J. Hu, G. Wang, F. L. J. tao Sun, Z. Chen, Understanding user's query intent with wikipedia, in: Proceedings of the 18th international conference on World wide web - WWW '09, 2009.

[25] C. T. Hemphill, J. J. Godfrey, G. R. Doddington, The atis spoken language systems pilot corpus, in: In Proceedings of the workshop on Speech and Natural Language - HLT '90, 1990, p. 96–101.

[26] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces, in: ArXiv, abs/1805.10190, 2018.

[27] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, I. Vulić, Efficient intent detection with dual sentence encoders, in: arXiv:2003.04807, 2020.

[28] S. Louvan, B. Magnini, Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, p. 480–496.

[29] R. Grishman, B. Sundheim, Message understanding conference- 6: A brief history, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, p. 466–471.

[30] E. Loper, S. Bird, Natural language processing with python: analyzing text with the natural language toolkit. (2009).

[31] J. Liu, P. Pasupat, S. Cyphers, J. R. Glass, Asgard: A portable architecture for multilingual dialogue systems, in: In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, 2013, p. 8386–8390.

[32] N. Jindal, B. Liu, Identifying comparative sentences in text documents, in: Proceedings of

the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 06, 2006, pp. 244–251.

[33] A. Panchenko, A. Bondarenkoy, M. Franzekz, M. Hageny, C. Biemann, Categorizing comparative sentences, in: In Proceedings of the the 6th Workshop on Argument Mining (ArgMining 2019), 2019.

[34] J. Bjerva, N. Bhutani, B. Golshan, W.-C. Tan, I. Augenstein, Subjqa: A dataset for subjectivity and review comprehension, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP, 2020, p. 5480–5494.

[35] H. Wachsmuth, M. Trenkmann, B. Stein, G. Engels, T. Palakarska, A review corpus for argumentation analysis, in: 15th International Conference on Intelligent Text Processing and Computational Linguistics, 2014, p. 115–127.

[36] S. Quarteroni, S. Manandhar, Designing an interactive open-domain question answering system, Natural Language Engineering 15 (2008) 73–95.

[37] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174. Klagenfurt, Germany: SSOAR.

[38] R. J. Moore, R. Arar, Conversational ux design: An introduction, Studies in Conversational UX Design (2018) 1–16. Springer International Publishing.

[39] K. Hamilton, S.-I. Shih, S. Mohammed, The development and validation of the rational and intuitive decision styles scale, Journal of Personality Assessment 98 (2016) 523–535.

[40] F. Costa, S. Ouyang, P. Dolog, A. Lawlor, Automatic generation of natural language explanations, in: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, 2018, p. 57:1–57:2.

[41] D. C. Hernandez-Bocanegra, T. Donkers, J. Ziegler, Effects of argumentative explanation types on the perception of review-based recommendations, in: Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), 2020.

[42] P. J. Nelson, Consumer information and advertising, in: Economics of Information, 1981, p. 42–77.

[43] L. Klein, Evaluating the potential of interactivemedia through a new lens: Search versus experience goods, in: Journal of Business Research, volume 41, 1998, p. 195–203.

Paper 6

The following paper is a manuscript version submitted to ACM:

- Hernandez-Bocanegra, D.C. & Ziegler, J. (2021). Explaining Recommendations Through Conversations - Argumentative Dialog Model and Comparison of Interaction Styles. Manuscript under review in ACM Journal Transactions on Interactive Intelligent Systems.

Explaining Recommendations Through Conversations - Argumentative Dialog Model and Comparison of Interaction Styles

DIANA C. HERNANDEZ-BOCANEGRA and JÜRGEN ZIEGLER, University of Duisburg-Essen, Germany

Providing explanations based on user reviews may facilitate users' assessment of a recommendation and improve the perception of the recommender system (RS) as a whole. While static explanations are dominant up to now, some interactive explanatory approaches have emerged in explainable artificial intelligence (XAI), making it easier for users to examine system decisions and to obtain more arguments supporting system assertions. However, little is known about how interactive interfaces should be conceptualized and designed to meet the explanatory aims of transparency, effectiveness and trust in RS. Thus, we investigate the potential of conversational explanations in review-based RS in the domain of hotels, and propose an explanation approach inspired by dialog models and formal argument structures. In particular, we address users' evaluation of a review-based RS which provides interactive explanations through two different interface types: GUI navigation, and natural language interface, the latter allowing users to express explanatory queries using natural language utterances.

Providing explanations by means of natural language conversation is a novel and promising way to support users' understanding of the items proposed by a RS. However, little is known about the impact and implications of providing such explanations, using for example a conversational agent (CA). Particularly, there is a lack of understanding of how users would formulate their questions in this context. In addition, there is also a lack of datasets that could facilitate the implementation of a dialog system for explanatory RS. Consequently, we describe in this paper the steps that led us to the implementation of a conversational agent for explainable RS, which involved the formulation of an intent model for explanatory queries in RS of the domain of hotels, and the building of ConvEx-DS [1], a dataset containing intent annotations of 1806 users' questions, which can be used to train intent detection procedures as part of the development of conversational agents for explainable RS.

Finally, we conducted a user study to compare users' evaluation of the two types of interactive explanations proposed, and to test the effect of varying degrees of interactivity that result in greater or lesser access to explanatory information. Our results show that providing interactive options for users to contest explanatory arguments of the system has a significant positive influence on the evaluation by users (compared to low interactive alternatives). Results also suggest that user characteristics such as decision-making style or visualization familiarity may have a significant influence on the evaluation of different types of interactive explanation interfaces.

CCS Concepts: • **Information systems → Recommender systems**; • **Human-centered computing → User studies**; **Natural language interfaces**.

Additional Key Words and Phrases: Recommender systems, explanations, argumentation, interactive interfaces, conversational agent, dataset, intent detection, user study

1 INTRODUCTION

Explaining the decisions or recommendations of artificial intelligence-based systems can yield several advantages for users' perception of these systems. Particularly, presenting explanations in recommender systems (RS) can contribute to a positive perception by users in terms of transparency, effectiveness, or trust [96]. In the past, many approaches to

[1] ConvEx-DS can be downloaded at https://github.com/intsys-ude/Datasets/tree/main/ConvEx-DS

Authors' address: Diana C. Hernandez-Bocanegra, diana.hernandez-bocanegra@uni-due.de; Jürgen Ziegler, juergen.ziegler@uni-due.de, University of Duisburg-Essen, Forsthausweg 2, Duisburg, Germany, 47057.

2022. Manuscript submitted to ACM

explaining the products or services suggested by an RS have been based on ratings assigned by customers with similar preferences, or on features of the recommended items, approaches related to collaborative and content-based filtering methods [40, 99]. More recently, and fueled by the advances in natural language processing, user-written reviews have received considerable attention as rich sources of information about an item's benefits and disadvantages. Here, the wealth of detailed information on the positive and negative aspects of the items can be exploited for generating both recommendations and explanations. However, the question of which review-based information to show and how to present it to impact positively users' perception remains largely open, mainly due a general lack of user-centric evaluations of explanatory RS [84].

Until now, the predominant approach in both RS, and explainable AI (XAI) in general, has been to provide explanations in a static manner, i.e. presenting explanations as a single unit of information, whether in numerical, textual or graphical form. The lack of interactivity in these explanation methods limits users' flexibility and control in scrutinizing the reasons of a recommendation according to their individual needs with respect to level of detail, different explanatory aspects, or form of presentation. In this paper, we therefore propose and investigate the concept of *interactive, conversational explanations* where users are free to explore explanatory information in a much less constrained manner. Whereas conversational RS have become a research topic of interest in recent years (see, e.g. [49]), accessing explanatory information in a conversational manner is a largely unexplored area, with a corresponding lack of empirical evidence for its potential effectiveness [94]. Conversational approaches may be realized in different forms, for example, as natural language-based dialogs with a chatbot, or as interaction steps in a graphical interface. We adopt here the interpretation of the term conversational as used in [49], defining conversational explanations as the provision of explanatory information in an interactive, multi-turn, dialogical process that may be instantiated as natural language dialog, GUI-based navigation, or other interaction styles.

Our work is grounded on the hypothesis that providing interactive options for examining explanatory information might positively impact users' evaluation of RS, by allowing the user to request, for example, an elaboration of the evidence for the claims made by the system, or, in contrast, statements questioning the appropriateness of a recommendation. User reviews are a rich and particularly relevant source for extracting such different types of explanatory information, but translating this information into effective user-system conversations is a largely open question. Consequently, we aim in this paper to answer the following research question:

RQ1 How can we leverage conversational explanations to increase the quality of review- based explanations in RS by users with respect to explanatory aims such as RS transparency and effectiveness, and trust in the RS?

When designing conversational explanations, the question arises as to how the dialog between user and system should be structured and which information components should be provided in the different interaction steps. Although a general theoretical model of explainable recommendations has not yet been established, we propose, as discussed in [45], to analyze conversational explanations through the lens of argumentation theory, which has resulted in a wide range of models of argumentation [6]. Most of these models define logical structures, with components like claims and the evidence to support or refute them. An alternative type of models departs from the idea of static argumentation, proposing instead a dialectical approach [104], with a focus on the exchange of arguments between two parties within a dialog. This concept provides the basis for dialog models of explanation as proposed by [71, 103], which take into account the social aspects of an explanatory process, as has been suggested by [46, 76]. This approach could facilitate the interactive provision of explanatory information, in the form of a question-and-answer exchange. However, the practical application of dialog models in explainable RS and their actual benefit from the users' perspective is yet to be determined.

Consequently, we address in this paper the concept of explanations as interactive argumentation, which we realized through two types of interface design: interactive navigation in a graphical user interface (GUI) and text-based natural language conversation in the form of a chatbot. In the former case, users can navigate through the explanatory information by asking questions such as "why-recommended" or "what was reported" on a certain aspect, using hyperlinks and buttons to this end. In contrast, by using the natural language conversation interface, users can type questions using their own words, within an interaction with a conversational agent. Thus, we compare users' evaluation of the RS when interacting with these two interface types, aiming to answer the following question:

RQ2: How do users evaluate review-based RS, when provided with interactive explanations through different interface types, namely *GUI-based interactive navigation* and *natural language conversation*?

To answer our research questions, we first formulate an explanation scheme, representing our conversational approach, aiming at facilitating the exploration of arguments that support the claims made by the system (e.g. an item is worth purchasing), while providing answers to users' explanation-related questions at different levels of detail, as described in section 3.

Next, we set out to investigate the unexplored implications and effects of conversational explanations in review-based RS, and in particular the use of conversational agents, given their ability to enable two-way natural language communication. This opens up the range of possible questions a user can ask the system, and supports the progression of explanations as a question-answer cycle, until the understanding of system claims is achieved by the explainee. The above could contribute to a better acceptance of explanations by users, as suggested by conceptual models of explanation based on dialog [46, 103].

Although user interfaces inspired by human-to-human conversation have been developed and used for a long time to assist users in a wide range of tasks [79], little is known about how a CA should be conceptualized or designed in the context of XAI, and in particular, in explainable RS. In the first instance, little is known about the type of explanation-related questions users would ask the RS. Although several datasets exist that support the development of dialog and question-answering (QA) systems, these are generally focused on open domain-search (e.g. [75, 90]), or specific processes such as flight or hotel booking, without a focus on explanatory interaction as such. In particular, and to our knowledge, there are as yet no publicly available datasets intended to support the development of an explanatory CA for RS. Specifically, there are as yet no datasets for detecting the user's intent expressed in a question in such a context. Therefore, as a first prerequisite to develop and evaluate such a dialog system, we investigated how users would formulate their questions to a CA in an explanatory context.

To this end, we performed a series of studies that are reported in Section 6. First, we conducted a Wizard of Oz (WoOz) pre-study [52] (section 6.1), that provided insights into the type of questions users might ask in an explanatory context. As a result, we formulate a dimension - based intent model, which we describe in detail in section 6.1.3. In order to check the validity of this proposal on a larger scale, i.e, the extent to which the model is able to accurately represent user intents given a larger set of questions, we performed a larger corpus collection, as described in Section 6.2. Here, we implemented a RS able to provide explanations in response to user generated questions. We also evaluated, as an indirect measure of validity, the evaluation of helpfulness of the responses by users, under the assumption that if the system has adequately recognized the user's intent, it is able to generate a response that approximates the user's information need, and thus be considered helpful. As a first major contribution, our studies resulted in a data set (ConvEx-DS) in which user-generated questions about recommendations are annotated with different types of explanation-related user intents based on a dimensional intent model we developed.

As a second major contribution, we developed ConvEx, a CA that includes a natural language understanding (NLU) module, based on classifiers trained on ConvEx-DS, using the state-of-the-art natural language processing (NLP) model BERT [26], and a dialog policy designed to enable conversation flows given different types of recognized intents. ConvEx allows not only to answer standalone questions, but also to provide follow-up arguments to support the system's responses to user requests.

As a further contribution, and to answer RQ2, we conducted a user study to compare users' evaluation of a RS under the two conditions: GUI-based navigation and natural language conversation with a CA, in terms of explanation quality, transparency, effectiveness, and trust. In addition, we also tested the effect of providing different degrees of interactivity to access explanatory information, as well as the mediating effects of psychological user characteristics, such as decision-making style and visualization familiarity.

We use the domain of hotels for our studies, since it represents an interesting mix between search goods (with attributes on which complete information can be found before purchase or use [83]) and experience goods (which cannot be fully known until purchase or use [83]). Such a product evaluation will likely benefit from third-party opinions [54, 83], potentially rich in argumentative information that can be used for explanatory purposes.

In this article, we extend work previously reported in [45] (GUI navigation, referred to in Sec. 5), in [43] (WoOz pre-study, reported in Sec. 6.1 and 6.5), and in [44] (dataset collection and annotation, addressed in Sec. 6.2 to 6.5. New content of this article addresses design and implementation details of the CA to provide natural language explanations (Sec. 6.6), and the final user study to compare the interfaces: GUI navigation and natural language conversation (Sec. 7).

Finally, the contributions of this paper can be summarized as follows:

- We introduce the novel concept of conversational explanations for recommender systems.
- We formulate a scheme for explanations as interactive argumentation in review-based RS, inspired by dialog models and argument structures. We also model and validate user intents for natural language explanatory queries in hotel domain RS.
- We release the dataset ConvEx-DS (**Conv**ersational **Ex**planations **D**ata**S**et), consisting of 1806 user questions with explanatory purpose in the domain of hotels, with question intent annotations, which can facilitate the development of explanatory dialog systems in RS.
- We implemented a conversational agent for explainable RS, ConvEx, which supports conversation flows for different types of intent, and enables users to access additional arguments that support system explanation attempts, at will.
- We provide empirical evidence of the effect that different types of interactive interfaces (GUI navigation and natural language conversation) have on users' evaluation of RS, as well as the effect of different degrees of interactivity to explore explanations, and the influence of user characteristics on such evaluation.

2 RELATED WORK

2.1 Explanations in RS

Providing explanations for recommended items can serve different purposes. Explanations may enable users to to understand the suitability of the recommendations, to understand why an item was recommended, or they may assist users in their decision making. Among the most popular approaches are the methods based on collaborative filtering (e.g. "your neighbors' ratings for this movie" Herlocker et al. [40]), as well as content-based methods that allow feature-based explanations, showing users how relevant item features match their preferences (e.g. Vig et al. [99]).

Review-based explanations. As summarized in [45], review-based explanatory methods leverage user generated content, rich in detailed evaluations on item features, which cannot be deduced from the general ratings, thus enabling the generation of more detailed explanations, compared to collaborative filtering (e.g. "Your neighbors' ratings for this movie" [40]) and content-based approaches (e.g. [99]). Review-based methods allow to provide: **1)** verbal summaries of reviews, using abstractive summarization from natural language generation (NLG) techniques [14, 22], **2)** a selection of helpful reviews (or excerpts) that might be relevant to the user, detected using deep learning techniques and attention mechanisms [17, 28], **3)** an account of the pros and cons of item features, usually using topic modelling or aspect-based sentiment analysis [27, 107, 111]. This information can also be integrated to RS algorithms like matrix or tensor factorization [5, 105, 111]) to generate both recommendations and aspect-based explanations.

Our work is based on the third approach, as reported in [45], and is particularly related to the Explicit Factor Model proposed by Zhang et al. [111], which integrates content extracted from reviews in a Matrix Factorization method. Based on this model, we can obtain statistical information on users' opinions (which has been proven to be useful for users [41, 81]), that can be provided in explanations through different presentation styles (verbal or visual). Moreover, we go beyond the basic model by examining interaction possibilities to explore the reasons behind the system's assertions at different levels of detail.

2.2 Interactive explanations

As discussed in [45], effects of interactivity have been studied widely in fields like online shopping and advertising [64, 95], and more specifically in the context of critique-based RS, where users are able to specify preferences and further criteria to refine the recommendations provided (e.g. [19, 66, 67]). Despite the intuitive advantages that interactivity can bring, it does not always translate into a more positive attitude towards the system, since it also depends on the context and the task performed [64]. Nevertheless, it has been shown that higher active control may be beneficial in situations where users have more information needs and a clearer mental goal [64], which also applies in our case (i.e. deciding which hotel to book).

Limited work in XAI has so far addressed interactive explanations, although to a much lesser extent compared to static explanations [1]. As discussed in [45], the dominant trend has been to provide mechanisms to check the influence that specific features, points or data segments have on the predictions of machine learning (ML) algorithms, as in [20, 57, 93]. However, the impact of such interactive approaches in explainable RS on the user remains largely unexplored. More specifically, the dominant ML interactive approach differs from ours in at least two ways: 1) we use non-categorical sources of information, subjective in nature and unstructured, which, however, can be used to generate both textual and visual structured arguments 2) such approach is designed to meet the needs of domain experts, i.e. users with prior knowledge of artificial intelligence (AI), while we aim to target the general public.

2.3 Conversational RS (CRS)

While a widely accepted definition of CRS has not yet been established, we adopt the following statement by Jannach et al. [49], "A CRS is a software system that supports its users in achieving recommendation-related goals through a multi-turn dialogue". This definition does not limit the way in which input is made by the user, so the term *conversational* does not exclude forms of interaction outside written or spoken text. Furthermore, and according to [49], input and output modalities of CRS, as found in CRS related literature, involve approaches based on:

1) Conventional web-based navigation, based on structured layouts and features, like buttons and hyperlinks. Throughout this article we will refer to this modality as *GUI navigation*. Particularly, we address our proposal of explanatory interface under this paradigm in Sec. 5.

2) Natural language (written or spoken).

3) Hybrid, a combination of natural language and other modalities. Under this approach, users can indicate their input using both natural language expressions and features such as buttons and other web controls. In section 6 we discuss our approach to explanations under this paradigm, which we will refer to as *natural language conversation* throughout the paper.

As also summarized by [49], CRS perform four main types of tasks: requesting user's preferences, recommending items, explaining decisions/predictions and responding to users' actions. However, as pointed out by [49], conversational explanations have been addressed to a much lesser extent, compared to the other CRS tasks, with only a few papers reported on this subject.

Our work differs from the existing approaches to CRS, as in work reported by [21, 110], who aimed to collect user preferences to generate recommendations through dialog in domains such as restaurants or electronics, usually neglecting explanations for recommended items, which is our central purpose. Our work also differs from conversational approaches in the hotel domain which focus on processes like customer service and booking assistance [13]. We aim, in contrast, to explore the implications and effects of using conversational interfaces to explain RS rationale. In one of the few works on the subject, [85] propose a model of social explanations for movie recommendations. However, according to their approach, it is the system that leads the conversation, providing justifications for recommendations even when they are not explicitly requested by the user, whereas according to our proposal, the user would have the active role, being enabled to ask the questions that lead to an argumentation by the system.

2.4 Dialog models of explanation

As discussed in [45], in contrast to static approaches to explanation, dialog models have been formulated conceptually [2, 71, 89, 101], allowing arguments over initial claims in explanations, and leading to an interactive exchange of statements. For example, Rago et al. [89] defined a protocol to provide conversational argumentative explanations in RS. They restrict, however,user interactions to a limited set of possible questions a user may ask, while we explore possibilities for users to express their explanatory needs in their own words.

This dialogical approach contrasts with other argumentative - though static - explanation approaches [4, 15, 41, 58, 109] based on static schemes of argumentation (e.g. [36, 97]), where little can be done to indicate to the system that the explanation has not been fully understood or accepted, and that additional information is still required, as initially discussed in [45]. Despite the potential benefit of using these models to increase users' understanding of intelligent systems [76, 106], their practical implementation in RS (and in XAI in general) still lacks sufficient empirical validation [71, 76, 94].

2.5 Question answering (QA)

Our work is closely related to QA systems which aim to answer questions posed by users in natural language, using information retrieval (IR) and NLP techniques, on various types of web documents or in knowledge bases, as initially discussed in [43]. While most QA systems are designed to reply to questions by offering excerpts from documents or lists of items, much less work has been devoted to advanced “how-to”, “why”, evaluative, comparative, and opinion questions [62, 77] that require usually the aggregation and comparison of multiple items over different pieces of information.

Lipton [63] defines *explanation* as an answer to a *why-question*, however, other types of questions can also be answered by explanations, i.e. how? what? [76], the latter being one that could be answered with a factoid sentence, for which we aim to support both factoid and advanced question types. Additionally, and in contrast to the common QA approach where the system replies to standalone questions, interactive QA involves a dialog interface enabling related, follow-up and clarification questions [88].

Our approach, as elaborated in [43], differs from most QA methods, especially those based on IR, because in our case responses are not generated solely on the basis of information sources, but should be consistent and reflect the mechanism used to generate the recommendations. Additionally, to answer complex questions (e.g., "why"), our approach focuses on the aspects most relevant for users, to provide concise and relevant statements, aggregating information from different reviews. To this end, our approach relies on the user profile inferred by the RS algorithm, especially when no explicit features are addressed in the user's question. Implicit user preferences are not taken into account in most QA approaches, which stems from their use of IR methods, where the relevance of a document is estimated based on how much its content is related to the query [78]. Additionally, we propose to follow an argumentative explanation structure to generate responses, which could improve users' evaluation of RS, as evidenced by using an interactive GUI navigation approach, and as we reported in [45]. Although argumentation has already been exploited in QA [80], its use has been mainly focused on the algorithmic aspects of information extraction, whereas very few approaches exploit argumentation as a way of presenting explanations in response to user queries [3].

2.6 Dialog systems and conversational agents

Often referred to as conversational agents, or dialog agents, dialog systems enable human-computer interaction by means of natural language expressions. Dialog systems can support interactions based on speech or written text. Such systems can be categorized as non-task-oriented (e.g. smalltalk systems), or task-oriented (e.g. booking or searching, or general assistants like Siri or Cortana) [18]. Our approach falls into the text-based, task-oriented group. Throughout this paper, we will refer to the terms dialog system and conversational agent interchangeably.

In [18], authors discuss two types of methods for implementing task-oriented dialog systems: pipeline and end-to-end methods. The first is the most widely applied, and is characterized by architectures that include the following components: language understanding (NLU modules to interpret intent and slots), a dialog state tracker (establishes the dialog state based on input and dialog history), dialog policy learning (to determine the next action in the dialog), and natural language response generation (NLG). On the other hand, end-to-end methods allow for joint training of all components. This approach can be beneficial in contexts where flexible adaptation is required, e.g. when the system is to be rapidly scaled to new domains or applications. On the other hand, the interdependency between pipeline components could make it difficult to adapt to new domains or data types (changes in one component may require changes in other components). Our implementation is framed within the pipeline approach, which remains beneficial for the early stages of dialog engineering in new domains [18], as in our case (explainable RS), and we leave for future work the exploration of end-to-end approaches.

Dialog management. The components of dialog state and dialog policy are usually grouped into a dialog management (DM) component. A taxonomy of DM approaches is discussed by [38], who differentiates between two main approaches: handcrafted (state and policy are defined as a set of rules defined by developers and domain experts) and probabilistic (rules are learned from corpora with real conversations). Our implemented system ConvEx falls into the former class, given that, although we compiled a dataset useful to detect user intention (ConvEx-DS), we still lack a corpus to facilitate the inference of states and sequences for conversations in the explanatory RS context.

Particularly, our implementation falls into the handcrafted approach known as finite-state, where the dialog state has a fixed set of possible transitions to other states. In our implementation, this was parameterized in a static file, where we set the sequence of possible actions to be executed by the system, given the recognized intent for a user query (see Section 6.6). More sophisticated approaches such as frame-based ones are used in commercial DM like Google's DialogFlow [2], which allows greater flexibility, e.g., adding a data model so that slots can be filled in any dialog sequence, which was not necessary for the purposes of this paper.

2.7 Mediation effect of user characteristics on users' evaluation of RS and explanations

It has been shown that a number of user characteristics may moderate the effect of interactive functionalities on the evaluation of explanations [64]. Regardless of its type, an explanation may not satisfy all possible explainees [94]. Moreover, individual user characteristics can influence the evaluation of a RS [55, 108], for which we assumed that this would also be the case for explanations, as discussed by [7, 41, 45, 56].

Since a main objective of providing explanations is to support users in their decision-making, investigating the effect of differences in user characteristics with respect to the decision-making process is of special interest to us. As discussed in [41, 45], decision-making styles are determined significantly by preferences and abilities to process available information [31]. Particularly, we focus on the moderating effect of the *rational* and *intuitive* decision making styles [37], the former characterized as a propensity to search for information and evaluate alternatives exhaustively, and the latter by a quick processing based mostly on hunches and feelings. We also considered *visualization familiarity*, i.e. the extent to which a user is familiar with graphical or tabular representations of information.

2.8 Further related work

In section 6, we discuss specific work related to: users' utterances on explanation needs, intent detection and slot filling, and datasets for intent detection.

3 CONCEPT: EXPLANATIONS AS INTERACTIVE ARGUMENTATION

To evaluate our research questions, we designed an interaction scheme for the exploration of explanatory arguments in review-based RS, as initially introduced in [45]. A recommendation issued by a RS can be consider a specific form of a claim, namely that the user will find the recommended item useful or pleasing [29, 45]. The role of an explanation is thus to provide supportive evidence (or rebuttals) for this claim [45]. Claims are, however, also present in the individual user's rating and opinions, which may require explaining their grounds as well, thus creating a complex multi-level argumentative structure in an explainable RS [45], a challenge also identified by [34]. To formulate an explanation scheme able to support this type of structure, and as an extension of work reported in [45], we considered dialog-based explanation models [71, 102, 103], in which instead of a single issue of explanatory utterances, an explanation is considered as an interactive, multi-turn process, where a user can indicate when additional arguments are required, to increase their understanding of system claims.

In this context, Walton proposes in [102, 103] a model which involves explanation requests (user questions) and attempts at explanation (namely, system responses referred to as assertions, with no specific structure). On the other

[2] https://cloud.google.com/dialogflow

hand, Madumal et al. [71] note that argumentation may occur within an explanation, and model the switch between an explanatory dialog and an argumentative one, specifying the explanatory loops that can be triggered as a consequence of follow-up questions. Allowed moves within an explanatory interaction can be defined on a basis of such dialog models. However, they offer little indication of how the arguments within the moves should be structured, so their acceptance by users can be increased. To this end, and as reported in [45], we based on the scheme by Habernal et al. [36], an adaptation of the Toulmin model of argumentation [97], formulated to better represent the kind of arguments usually found in user-generated content. The scheme by [36] includes the following components: claim (argument conclusion), premise (overall reason to accept the claim), backing (additional specific evidence supporting the claim), rebuttal (statement attacking the claim, representing an "opposing view") and refutation (statement attacking the rebuttal).

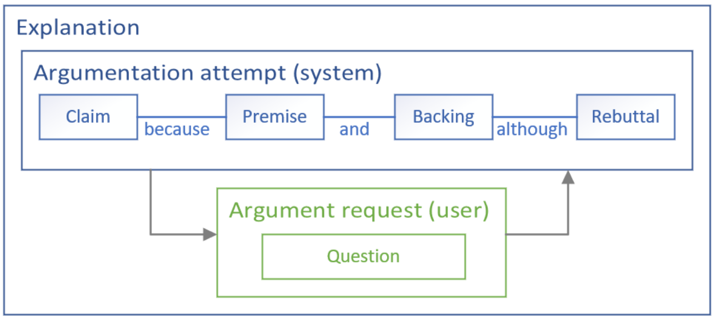


Fig. 1. Overall scheme for explanations as interactive argumentation in review-based RS. Argumentation attempts (in blue) are constituted by different argument components (claim, premise, etc.)

Our proposed overall scheme is shown in Figure 1. Unlike Walton, who modeled explanatory movements as explanation requests and attempts, we considered an explanation as an iterative process of *argumentation attempts* (the system intends to provide arguments to explain something) followed by *argument requests* (the user asks the system to provide - follow-up - arguments that support the claim that user will find the recommended item useful) [45]. With this slight renaming of the explanation components, we aim to make explicit that, unlike Walton's model, statements within system's attempts at explanation follow an argument structure. Thus, in our approach, argumentation attempts include premises (overall reason to accept the claim that a recommended item is worthy to be chosen, e.g. "Good food, great staff"), backing (e.g. percentage of positive opinions about an aspect, or actual customer comments supporting the aggregate percentages), and rebuttal (e.g. percentage of negative opinions about an aspect, or the actual negative comments by customers) [45].

We operationalized our proposal by implementing a RS using two types of interfaces or interactive paradigms:

- GUI navigation: Here, design features like links and buttons enable argument requests by users, while argumentation attempts involve features like text, tables and bar-charts (for aggregated opinions) and lists (for excerpts of customer reviews). In section Sec. 5 we unfold the overall explanation scheme, to formulate the interaction flow design, and the layout of the interface design features.
- Natural language conversation: Here, argument requests can be indicated by users by typing questions in their own words, or using buttons and dropdown menus for quick utterances (e.g. to indicate “yes”/”no”, “positive”/”negative”, or a hotel aspect), while argumentation attempts involve text (for aggregated opinions, and

other answers to user questions) and lists (for excerpts of customer reviews). In this case, the overall explanation scheme serves as a basis for the CA dialog policy. We elaborate further design details for this type of interface in Sec. 6.

Our proposal of explanations as interactive argumentation fosters the interactive features defined by [64]: active control (voluntary actions that can influence the user experience) and two-way communication (the ability of two parties to communicate with each other). These features are reflected in our proposal by the possibility to examine explanatory information at will, e.g. to access excerpts from customer reviews supporting system claims (e.g., an item is worth choosing), and to indicate to the system which aspects are of greater relevance to the user, so that content is presented accordingly. Consequently, our proposal opposes a single static view that may include aspects or details that are not relevant, or leave out those that are.

We describe in Sec. 5 further details of the GUI navigation design, and in Sec.6 the steps we performed to design and implement a CA representative of the natural language interface. Finally, in section 7 we report results of a user study comparing both types of interfaces. We also tested in this user study the effects of providing different degrees of interactivity, namely low (users only have access to an overall explanation), and high (users can access additional information that supports argumentation attempts by the system, e.g. excerpts of customer comments that form aggregated opinions, filtered by specific aspect or feature).

4 BASIC COMPONENTS OF EXPLAINABLE REVIEW-BASED RS

User and data collection studies reported in this article were conducted using a review-based RS, implemented to generate both recommendations and explanations, according to our proposed explanation scheme, and under different types of interface. Implementation details of this base component, which were originally reported in [45], are described below.

4.1 Dataset for aspect-based sentiment detection

We used the ArguAna dataset [100] for the detection of the aspect addressed in a statement, and the sentiment expressed about it. ArguAna includes hotel reviews and ratings from TripAdvisor, where sentiment and explicit features are annotated sentence wise. We further categorized the explicit features in 10 general features (room, price, staff, location, facilities, bathroom, ambience, food and beverages, comfort and checking), with the help of 2 annotators (Krippendorff's alpha of 0.72), as reported in [45], aiming to train a classifier to detect the main aspect addressed in a sentence (e.g. "I loved the bedding" would be classified as *room*).

4.2 Explainable RS method

We implemented the Explicit Factor Model (EFM) [111], a review-based matrix factorization (MF) method to generate both recommendations and explanations, as reported in [45]. The rating matrix (ratings granted by users to items) consisted of 1284 items and 884 users extracted from the ArguAna dataset (only users with at least 5 written reviews were included), for a total of 5210 ratings. Item quality and user preference matrices were calculated using the sentiment detection described previously. Each element of the item quality matrix contains the measurement of the quality of the item for each aspect, while the elements of the latter measure the extent to which the user cares about an aspect. The number of explicit features was set to 10. Model-specific hyperparameters were selected via grid-search-like optimization. After 100 iterations, we reached an RMSE of 1.27, a metric used to measure the differences between dataset values and the values predicted by the RS model. Predicted ratings were used both to sort recommendations, and are shown within explanations (average hotel rating with 1-5 green circles shown in recommendations list). Values of quality matrix were used to calculate the percentages of positive and negative comments on aspects.

4.3 Aspect-based sentiment detection

We trained a BERT classifier [26] to detect the general feature addressed within a sentence: we used a 12-layer model (*BertForSequenceClassification*), 6274 training sentences, 1569 test sentences, F-score 0.84 (macro avg.), as reported in [45]. We also trained a BERT classifier to detect the sentiment polarity, using a 12-layer model (*BertForSequenceClassification*), 22674 training sentences, 5632 test sentences, *F*-score 0.94 (macro avg.). The classifier was used to **1)** obtain the quality of hotels (item quality matrix in EFM) and relevance of aspects to users (user preference matrix in EFM), and **2)** to present participants with negative and positive excerpts from reviews regarding a chosen feature during the explanatory interaction, as elaborated in [45].

4.4 Personalization mechanism

To overcome implications of the *cold start* problem [91] (system does not have enough information about the user to generate an adequate profile and thus, personalized recommendations), participants were asked beforehand (as part of the respective user study) for the five hotel features that mattered most to them, in order of importance. The system calculated a similarity measure to detect users in the EFM preference matrix with similar preferences. Next, the user in the matrix with the preferences most similar to those of the current user was used as a proxy to generate recommendations, i.e. we selected the proxy user's predicted ratings, and use them to sort the recommendations and features within the explanations that were shown to the current user.

5 GUI NAVIGATION

We implemented a user interface under the GUI navigation paradigm, grounded on the base RS described in Sec. 4. We first unfolded the explanation scheme addressed in 3, as shown in Figure 2, to design the explanatory interaction flow. Consequently, the realization of the scheme as an user interface is depicted in Figure 3, as originally reported in [45].

Here, design features like links and buttons enable argument requests by users, e.g: the link "what was reported?" (Fig. 3b) fosters the interactive features *active control* (control on what aspects the system should focus on in its argumentation) and *two-way communication* (user communicates to the system that further argument backing is needed), which triggers a system argumentation attempt.

As discussed in [45], an explanatory dialog can take place both through verbal interactions and through a visual interface (non-verbal communication, or a combination of verbal and visual elements) [71, 76]. As for presentation, while arguments are usually associated with oral or written speech, arguments can also be communicated using visual representations (e.g. graphics or images) [9]. Thus, the following display styles for the argumentation attempt "% of positive and negative opinions" were considered in [45]: table, bar chart, and only text. Results of a user study (n=170) were reported in [45], in which we aimed to test the effect of display style, as well as of different degrees of interactivity (high and low), as depicted in Fig. 4. While a significant effect of degree of interactivity was found (more interactive explanations contributed to a more positive evaluation of explanation quality, and of effectiveness and trust in the system), no main effect of display style was found.

We used the GUI navigation implementation reported in [45], for the comparison with the natural language interface (Sec. 6), as part of the user study reported in Sec. 7. To this end, we opted for the table display style to provide the

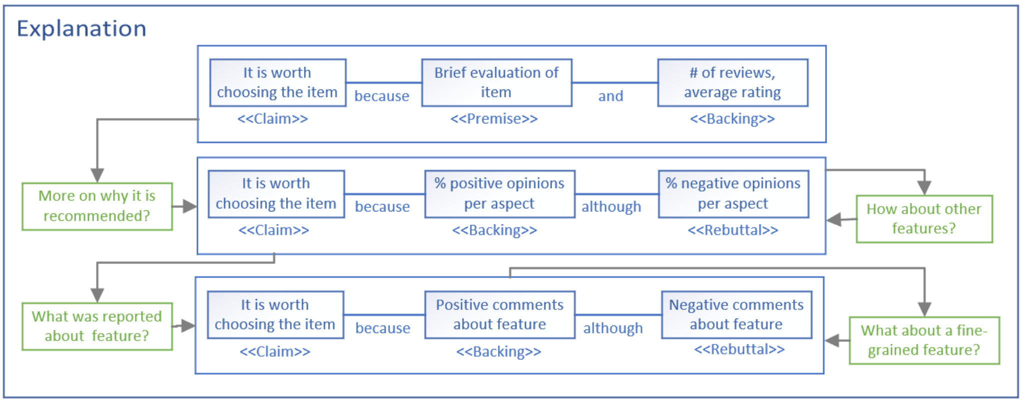
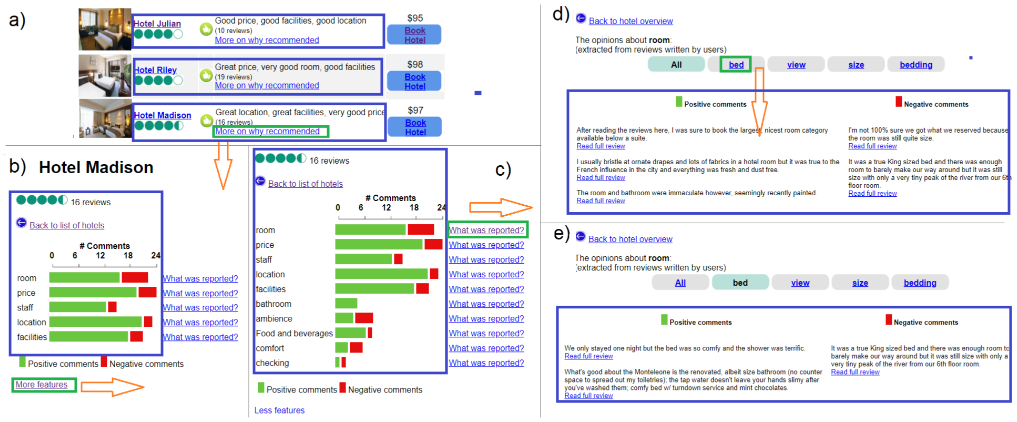

Explanation
It is worth choosing the item
<<Claim>>
because
Brief evaluation of item
<<Premise>>
and
# of reviews, average rating
<<Backing>>
More on why it is recommended?
It is worth choosing the item
<<Claim>>
because
% positive opinions per aspect
<<Backing>>
although
% negative opinions per aspect
<<Rebuttal>>
How about other features?
What was reported about feature?
It is worth choosing the item
<<Claim>>
because
Positive comments about feature
<<Backing>>
although
Negative comments about feature
<<Rebuttal>>
What about a fine-grained feature?

Fig. 2. Unfolded scheme for explanations as interactive argumentation in review-based RS, reported in [45]. Blue boxes: argumentation attempts by the system, green boxes: argument requests by users.



Hotel Julian
Good price, good facilities, good location
(10 reviews)
More on why recommended
$95

Hotel Riley
Great price, very good room, good facilities
(19 reviews)
More on why recommended
$98

Hotel Madison
Great location, great facilities, very good price
(16 reviews)
More on why recommended
$97

Hotel Madison
16 reviews
Back to list of hotels

room	What was reported?
price	What was reported?
staff	What was reported?
location	What was reported?
facilities	What was reported?

Positive comments Negative comments
More features

16 reviews
Back to list of hotels

room	What was reported?
price	What was reported?
staff	What was reported?
location	What was reported?
facilities	What was reported?
bathroom	What was reported?
ambience	What was reported?
Food and beverages	What was reported?
comfort	What was reported?
checking	What was reported?

Positive comments Negative comments
Less features

Back to hotel overview
The opinions about **room**:
(extracted from reviews written by users)

Positive comments

After reading the reviews here, I was sure to book the largest, nicest room category available below a suite.
Read full review

I usually bristle at ornate drapes and lots of fabrics in a hotel room but it was true to the French influence in the city and everything was fresh and dust free.
Read full review

The room and bathroom were immaculate however, seemingly recently painted.
Read full review

Negative comments

I'm not 100% sure we got what we reserved because the room was still quite size.
Read full review

It was a true King sized bed and there was enough room to barely make our way around but it was still size with only a very tiny peak of the river from our 6th floor room.
Read full review

Back to hotel overview
The opinions about **room**:
(extracted from reviews written by users)

Positive comments

We only stayed one night but the bed was so comfy and the shower was terrific.
Read full review

What's good about the Monteleone is the renovated, albeit size bathroom (no counter space to spread out my toiletries); the tap water doesn't leave your hands slimy after you've washed them; comfy bed w/ turndown service and mint chocolates.
Read full review

Negative comments

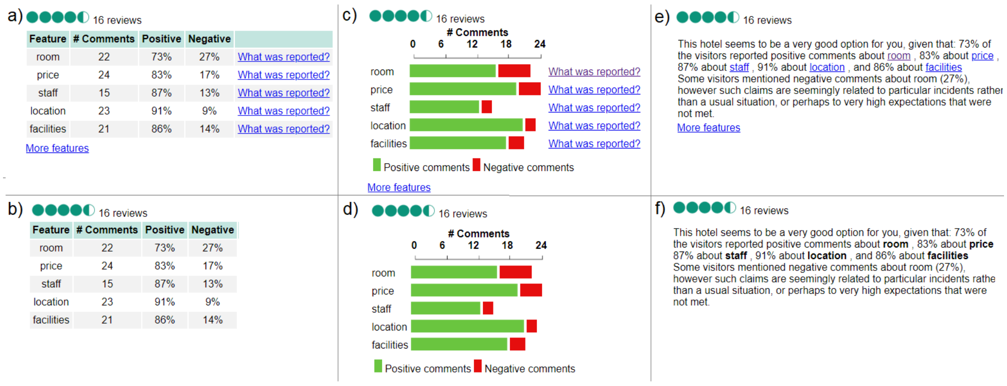
It was a true King sized bed and there was enough room to barely make our way around but it was still size with only a very tiny peak of the river from our 6th floor room.
Read full review

Fig. 3. Interactive explanations through a GUI navigation interface (screenshots of implemented system, reported in [45]). Enclosed in blue: argumentation attempts; in green: argument requests. Orange arrows: sequence of allowed moves, pointing to the next interface. a) List of recommended items; clicking on "More on why recommended" displays: b-c) aggregation of comments by aspect; clicking on "What was reported?" displays: d) comments on chosen aspect; clicking a fine-grained feature button, displays: e) comments on chosen feature.

aggregate opinion view, as it represents a middle ground between the graph (bar chart) and text-based representations mentioned above.

6 CONVERSATIONAL AGENT FOR EXPLAINABLE REVIEW-BASED RS

In this section we present our methodological approach to providing explanations as natural language conversation, and describe the steps we took for the implementation of a CA that provides such explanations. First, we report the results of a WoZ pre-study [43], which aimed to address the question of how RS users communicate their explanation needs using a CA. We then report details on the collection and annotation of ConvEx-DS [44] (a dataset for the automatic detection of question intent with explanatory purpose, as a prerequisite for the development of our CA). During corpus



a) ●●●●◐ 16 reviews

Feature	# Comments	Positive	Negative	
room	22	73%	27%	What was reported?
price	24	83%	17%	What was reported?
staff	15	87%	13%	What was reported?
location	23	91%	9%	What was reported?
facilities	21	86%	14%	What was reported?

More features

b) ●●●●◐ 16 reviews

Feature	# Comments	Positive	Negative
room	22	73%	27%
price	24	83%	17%
staff	15	87%	13%
location	23	91%	9%
facilities	21	86%	14%

c) ●●●●◐ 16 reviews

d) ●●●●◐ 16 reviews

e) ●●●●◐ 16 reviews

This hotel seems to be a very good option for you, given that: 73% of the visitors reported positive comments about room , 83% about price , 87% about staff , 91% about location , and 86% about facilities
Some visitors mentioned negative comments about room (27%), however such claims are seemingly related to particular incidents rather than a usual situation, or perhaps to very high expectations that were not met.

More features

f) ●●●●◐ 16 reviews

This hotel seems to be a very good option for you, given that: 73% of the visitors reported positive comments about **room** , 83% about **price** 87% about **staff** , 91% about **location** , and 86% about **facilities**
Some visitors mentioned negative comments about room (27%), however such claims are seemingly related to particular incidents rathe than a usual situation, or perhaps to very high expectations that were not met.

Fig. 4. Manipulation of *presentation style* in combination with *interactivity degree*, in user study reported in [45]. Left: *table*, middle *bar chart*, right *text*. Top: interactivity *high*, bottom: interactivity *low*.

collection, we also measured users' evaluation of the helpfulness of the answers that the system generated automatically to their questions, results that are also reported in this section. Finally, we describe the design and implementation details of ConvEx, a CA for explainable review-based RS, which we used for the user evaluation reported in section 7, under the natural language interface condition.

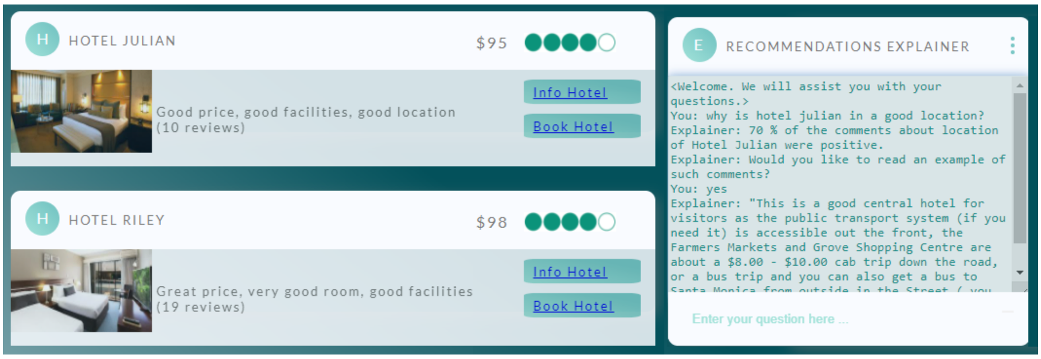
6.1 Wizard of Oz pre-study

In this section, we describe the details of the pre-study that led to the formulation of our dimension-based intent model. The content of this section was initially reported in [43].

6.1.1 Background. The design of adequate conversational explanations requires a proper understanding of possible user requests [43, 61]. [60] collected a dataset consisting of written conversations between humans with a movie recommendation goal, however, no explanatory inquiries like "Why do you recommend X?" are addressed. [8] collected a QA dataset for several domains (including hotels), which can be used to generate answers to factoid and subjective questions (e.g. "How is it the location?"), in regard to a specific item, but leaving out comparison and why-recommended queries. On the other hand, [61] formulated a XAI question bank, with questions users might typically ask about AI algorithms, but targeting users with expert knowledge in AI, whereas no similar question bank definition has been developed, to our knowledge, for end users and, in particular, for RS. Consequently, we first conducted a pre-study using the WoOz [52] method, to capture the possible questions users would ask in the context of RS explanations, particularly in the hotel domain, and a subsequent online study to collect a larger number of user-generated questions.

6.1.2 Methods. WoOz is a technique that has been widely adopted for human computer interaction prototyping [25, 72], in which a member of the research team (the *wizard*) simulates the response actions of the system, through a computer-mediated interface. We aimed, with this pre-study, to explore how users would interact with a conversational interface for explanations in RS. Here, we used as a basis the RS described in 4, and a dialog guideline for the wizard (see Appendix A.1, Fig. 15), based on the concept of explanations as interactive argumentation (Sec. 3), aiming to achieve a structured conversation similar across participants. We recruited 20 participants through Amazon Mechanical Turk, and collected a total of 20 dialogues and 105 user utterances. Participants were requested to interact with a RS, and to type any question of interest about one or more hotels in the chat box located on the right of the hotel list (Figure 5); we underlined that the chat box was designed to explain the reasons for the recommendations, to prevent the user from asking questions about other processes, such as booking assistance.

In respect of data analysis, we follow an inductive category formation [73]. This step involved two independent annotators, who came to a Cohen's kappa = 0.91, almost perfect agreement intercoder reliability [59]. To analyse follow-up questions, we validated anaphoras ("a linguistic form whose full meaning can only be recovered by reference to the context" [88]) and ellipsis ("an omission of part of the sentence, resulting in a sentence with no verbal phrase" [88]) , using criteria from [88] and [10]. See [43] for further details on the WoOz setup, participants demographics and payment, and subsequent data analysis.



Good price, good facilities, good location
(10 reviews)

Great price, very good room, good facilities
(19 reviews)

<Welcome. We will assist you with your questions.>
You: why is hotel julian in a good location?
Explainer: 70 % of the comments about location of Hotel Julian were positive.
Explainer: Would you like to read an example of such comments?
You: yes
Explainer: "This is a good central hotel for visitors as the public transport system (if you need it) is accessible out the front, the Farmers Markets and Grove Shopping Centre are about a $8.00 - $10.00 cab trip down the road, or a bus trip and you can also get a bus to Santa Monica from outside in the Street (you

Fig. 5. User interface presented to participants in WoOz pre-study [43].

6.1.3 Results - Dimension-based intent model.

Intents and entities. We found that users' intents could be classified into two main types: *domain-related* intents (regarding hotels and their features), and *system-related* intents (regarding the algorithm, or the system input) [43]. In turn, domain-related intents could be categorized according to the following dimensions [43]:

- *Scope*: Whether the question refers to a single item (*single*), a limited list of items (*tuple*), or to no particular item (*indefinite*).
- *Comparison*: Whether the question is (*comparative*) or not (*non-comparative*). We adopt the comparative sentence definition by [51] "expresses a relation based on similarities or differences of more than one object", including superlatives and relations like "greater" or "less than".
- *Assessment*: Whether the question refers to the existence or characteristics of item features (*factoid*), to a subjective assessment of the item or its features (*evaluation*), or to system reasons to recommend an item (*why-recommended*).
- *Detail*: Whether the question inquires for a specific aspect or feature (*aspect*), or for the overall item (*overall*).

Consequently, the intent of a single domain question could be defined as a combination of the 4 dimensions. Table 1 provides intent examples, their frequency in collected utterances, and the suggestion of possible answers that the system could provide. All but one of the questions could be classified as system-related intent, namely: "why are there so few reviews?". All questions of domain intent regarded the entities: *hotel* and *hotel feature*.

Table 1. Most frequent domain intents (combination of dimensions values) sorted by number of questions per intent (desc.), as reported in [43]

Scope	Comparison	Assessment	Detail	Example	# Qs	Type of initial response
Single	Non-comp	Factoid	Aspect	Does Hotel Julian have a pool?	29	Y/N or value
Single	Non-comp	Why-recomm	Overall	Why is Hotel Julian my top recommendation?	14	Because [Argument backing]
Single	Non-comp	Evaluation	Aspect	How is the food at Hotel Evelyn?	8	[Argument claim], because [Argument backing]
Indefinite	Comparative	Evaluation	Aspect	Which hotel has the best customer service?	7	Hotel X, because [Argument backing]
Indefinite	Non-comp	Factoid	Aspect	Do any of the hotels provide free breakfast?	6	Y/N or value
Tuple	Non-comp	Factoid	Aspect	what are the checking in times for hotel owen and hotel evelyn?	4	Y/N or value
Indefinite	Comparative	Evaluation	Overall	Which hotel has the best reviews?	4	Hotel X, because [Argument backing]
Indefinite	Non-comp	Evaluation	Aspect	what rooms would be good for parents with children?	3	Hotel X, because [Argument backing]
Tuple	Comparative	Evaluation	Overall	What is difference between hotel evelyn and hotel james?	2	Hotel X has better comments on [feature x] and [feature y].

Follow-up questions. We found that 55% could be classified as standalone questions and 26% as follow-up questions. A special case are inquiries that could work as both types (19%). Such is the case for comparative questions under the value "Indefinite" of dimension *scope*, which may refer to the best among all possible options (e.g. "which is the best hotel?") or, if a subset of options was previously discussed, as a follow-up, (e.g. "I am choosing between the Riley and the Evelyn. Which is the best hotel overall?").

In regard to the follow-up question types: 24% corresponded to pronoun or possessive adjective anaphora (e.g. " I'm looking for facts about current internet service - is *it* unchanged or upgraded?"), 71% noun phrase anaphora (e.g. "When was the last time *the Hotel* underwent a remodel?), and 5% ellipsis (e.g. "what are the checking in times for hotel owen? *and hotel evelyn?*"). We noted that pronouns and noun phrases in anaphora referred only to hotels names or hotel features. Moreover, 11% of utterances contained non-question sentences aiming to establish a context for a subsequent question, e.g. "I like the ambiance of the Hotel Evelyn, how were the reviews for that?". Finally, only 2.4% of utterances contained more than one question.

6.2 Collecting a corpus for intent detection: Background and Methods

In this section we describe the details of the data collection that led to ConvEx-DS consolidation. The content of this section was initially reported in [44].

6.2.1 Background.

Intent detection and slot filling. Intent detection seeks to interpret the user's information need expressed through a query, while slot filling aims to detect which entities - and also features of an entity - the query refers to [44]. In the open-search domain, intent classification is the most common approach for intent representation[98]. Here, an user query can be categorized using a classification scheme: a set of dimensions or categories [12, 98]. Our proposed intent model [43] corresponds to this representation type.

As discussed in [44], a large body of previous work has addressed the task of intent detection, both for open search domains (see [47, 98]) and task-oriented dialog systems, for processes such as flight booking, music search or e-banking, e.g. [16, 23, 39]. Methods proposed to solve these tasks range from conventional text classification methods, to more complex neural approaches, based on recurrent neural networks, attention-based mechanisms and transfer learning (we refer readers to the survey on this matter by Louvan and Magnini [69]).

Our work is related to the text classification approach, leveraging the representation of intents according to a classification scheme. Thus, the complex intent detection task can be split into small text classification tasks, to detect the class that best represents a sentence, for each dimension. To this aim, we leverage the language model BERT [26]. We solve the slot-filling task as a named entity recognition (NER) task as in [35], to detect entities referred in sentences (i.e. hotel names), for which we used the natural language toolkit NLTK [68].

Datasets for Intent detection. Given that, to our knowledge, there were no datasets to support the intent detection task in the specific context of explainable RS, we analyzed datasets that could support text classification according to the different dimensions of the intent model:

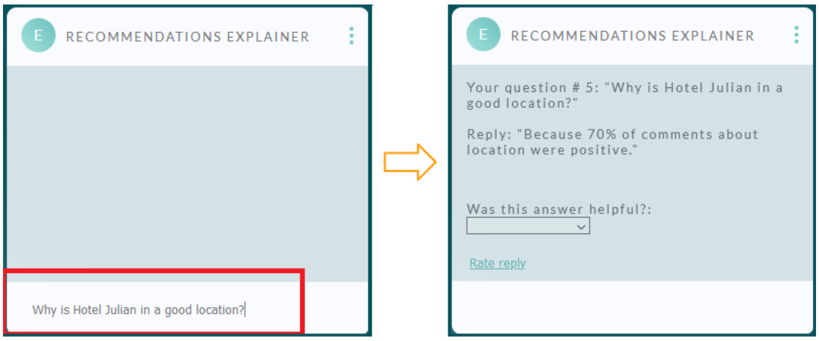
- Dimension comparison: Jindal and Liu [51] released a dataset for classification of comparative and non-comparative sentences, using news articles, reviews and forums on electronics and diverse brands, as a source.
- Dimension assessment: Bjerva et al. [8] released SubjQA, to detect subjectivity of questions in QA tasks, for different domains, including hotels. This dataset does not address *why-recommended* questions.
- Dimension detail: To our knowledge, no dataset addresses explicit annotation of sentences addressing overall item quality (e.g., "how good is Hotel x?"), in contrast to aspect-based sentences (e.g. "how good is the food").
- Dimension scope: Values of this dimension can be inferred from entity detection (e.g. number of entities referred), for which NER can be used.

6.2.2 Methods.

System for data collection. The RS described in Section 4 was extended by adding a NLU module to detect users' intent, using text classifiers trained initially on the auxiliary datasets described above ([8, 51], and data collected in [43] for dimension detail), as well as a module to generate answers consistent with the method on which the RS is based (EFM [111]). Development and corpus collection involved a process consisting of several iterations, as reported in detail in [43], aiming to refine methods for answer generation. For example, early iterations included methods to respond to the most common user intents; new iterations included responses to new occurrences of intents observed in the previous iteration.

Intent detection. As reported in [44], we trained BERT classifiers [26], one for each dimension (comparison, assessment and detail), using a 12-layer model (*BertForSequenceClassification*, bert-base-uncased), batch size 32, and Adam optimizer (learning rate = 2e-5, epsilon = 1e-8). Classifiers for comparison and assessment converged after 4 epochs, while for detail 5 epochs were needed. Datasets were split randomly into training (80%) and test (20%) during the training phase. In order to avoid overfitting, the most represented class was downsampled (randomly) to approximate the size of the less represented class, which was slightly upsampled (randomly) to fit round numbers like 1000 and 500. In the case of the detail dimension, due to the small size of the auxiliary dataset, both classes were increased to 100 instances each (using procedure described in [44]). Datasets (original and balanced) sizes are reported in table 2.

Answer generation module. We implemented a module to generate replies based on the intent detected. According to this proposal, *factoid* questions could be replied with Y/N or a value (e.g. check in times) based on metadata. As for *evaluation* or *why-recommended* questions, replies were based on the aggregation of positive or negative opinions regarding an aspect (if question was *aspect* based), or the most important aspects for the participant (in case question was *overall*). If the question was comparative, the system calculated which hotel was better among a tuple, or the best in general (scope *indefinite*), based on the aggregation of the opinions. These are some examples of the type of responses generated by the system: Q: "Why does Hotel Hannah have the highest rating?", A: "Because of the positive comments reported regarding the aspects that matter most to you: 86% about location, and 85% about price."; Q: "Which hotel is best, Hotel Lily, Hotel Amelia or Hotel Hannah?", A: "Hotel Lily has better comments on the aspects that are most important to you (location, facilities, staff). However, Hotel Amelia has better comments about room, price."; Q: "Hotel Amelia is described as having a great room, what makes it great?, A: "Comments about rooms are mostly positive (90%).".


E
RECOMMENDATIONS EXPLAINER
E
RECOMMENDATIONS EXPLAINER
Was this answer helpful?:
Rate reply

Why is Hotel Julian in a good location?

Your question # 5: "Why is Hotel Julian in a good location?"

Reply: "Because 70% of comments about location were positive."

Fig. 6. User interface for corpus collection, as reported in [44]. Box to write questions in red. Then, system shows answer and requests user to rate helpfulness. (Recommendations list was displayed on the left side as in Fig. 5)

6.2.3 Corpus collection study. 298 participants were recruited through the crowdsourcing platform Prolific, who interacted with the RS, and typed their questions about the reasons why the hotels were recommended, using their own words. We stated in the instructions that "The aim of the system is to provide explanations based on your questions" (to prevent the user from asking questions about other processes, such as booking assistance). For each question, and once the system response was displayed, users were asked to rate the helpfulness of the answer. Further details on study setup, participants demographics and responses quality check are reported in [44]. A total of 1836 questions users' written questions were collected in this step.

6.3 Corpus annotation

In this section we describe the details of the data annotation that led to definition of the gold standard for ConvEx-DS. The content of this section was initially reported in [44].

6.3.1 Intent type annotation. Sentences where first classified according to the types: domain-related intents (regarding hotels and their features), and system-related intents (regarding the algorithm, the system Input, or system functionalities). Given that we detected that domain questions clearly outnumbered system questions in our WoOz study, research

Table 2. Inter-rater reliability of ConvEx-DS, as reported in [44]. *Fleiss' kappa* refers to each dimension, and the *% of full agreement* to each class, i.e. percentage of questions in which all annotators agreed on assigning that class).

Dimension	Fleiss' kappa	Class	% of full agreement
Comparison	0.72	Comparative	77.28%
		Non-comparative	86.86%
Assessment	0.65	Factoid	73.99%
		Evaluation	58.56%
		Why-recommended	66.42%
Detail	0.75	Aspect	95.73%
		Overall	66.82%

team members annotated this class (98.3% agreement), as the low number of system-related questions could lead to the category being ignored in the crowdsourcing setup. Disagreements were resolved in joint meetings.

6.3.2 Dimension-based annotation. Domain-related sentences were used for the dimension-based annotation. We collected annotations for comparison, assessment and detail as independent tasks. The dimension scope was not annotated under the proposed procedure (is not a classification task but a NER task).

Annotators and crowdsourcing setup. Every sentence was annotated by 3 annotators: one member of the research team, and the other two crowdsourced on the Prolific platform. We divided the set of questions into 19 blocks of 100 sentences each, and every block had to be annotated for each dimension separately. The annotator from research team annotated all blocks for the three dimensions, while different crowdsourcing workers could annotate different blocks for different dimensions (same annotator did not annotate the same block for the same dimension more than once). Each block included 5 attention checks, and annotation guidelines remained visible during the task (see Appendix A.2). A total of 92 Prolific workers performed the task, but only responses of 57 workers were used for the calculation of Inter-rater reliability and the deduction of gold standard, after applying the quality check. Further details on study setup, participants demographics, payment, responses quality check and processing are reported in [44].

6.4 Corpus collection and annotation: Results

We report in this section the annotation statistics, and results of the validation of the intent model, which included: evaluation of helpfulness of answers to user questions, and evaluation of the performance of classifiers trained and tested on ConvEx-DS. The content of this section was initially reported in [44].

6.4.1 Annotation statistics. 30 out of 1836 collected questions were discarded (nonsense statements, or highly ungrammatical), for a final set of 1806 of annotated questions. Of these, only 24 were annotated as system-related questions. Length of questions: characters M=39.2, SD=15.67, words M=7.35, SD=2.86. We found a Fleiss' kappa of 0.72 for *comparison*, of 0.65 for *assessment* and of 0.75 for *detail*, indicating a "substantial agreement" [59], for all three dimensions. As for classes with lower percentages of questions with full agreement, we identified the following main causes:

Dimension assessment. - *Why-recommended* questions rated as subjective, given that adjectives like 'good' or 'great' are included in sentences, e.g. "why is hotel hannah location great?". - Questions that should be replied with a fact, but include adjectives that indicate subjectivity, e.g. "does hotel emily have any bad reviews?", "are there good transport links?", "which hotel best fits my needs?". - Questions with adjectives as 'cheap', 'expensive', 'close', 'near', 'far', which can be answered with either subjective or factoid responses, e.g. "Which is the cheapest hotel?", "is there an airport near any of these hotels?”. - Questions of the type "what is ... like", e.g. “What is the room quality like at Hotel Emily?” (this type of questions were actually not addressed in instructions).

Dimension detail. - Concepts that were regarded as hotel aspects, e.g. value (Which is best value for money), ratings (What is the highest rating for Hotel Levi), reviews (Which hotel has the most reviews?), stars (5 stars hotels?).

Finally, we have detected some questions that could hardly fit in the planned classes, e.g. “How do you define expensive? Do you compare against facilities and what is included in the price?”, “The Evelyn has 17 reviews and a positive feedback but scores lower than others with less reviews. Why is this?”. However, we found this number to be rather low (16 questions).

6.4.2 Helpfulness of system answers. Taking into account its the most refined versions, the system was able to generate an answer in 80.58% of the cases, and to partially recognize the intent or entities in 7.34% of the cases (thus asking the user to rephrase or indicate further information). Among the main reasons why questions could not be answered we found: complexity of the question or no information available to reply (31%), text that could be improved when replying to factoid questions (23%), wrong intent classification (11%), and system errors (11%).

Figure 7 (middle) shows the evaluation of answers’ helpfulness, according to ratings granted by study participants, across all iterations (M=3.58, SD=1.34). When taking into account only the last two iterations (which account for 63.85% of sentences collected, and involve the most refined versions of the system), we observed a greater helpfulness (to M=3.70, SD=1.30). We considered as "non-helpful" responses those that were marked with the values "Strongly Disagree" and "Disagree" when participants were asked "Was the answer helpful?". We analyzed the responses given to those questions by the system and found that in 34% of cases, replies provided actually make sense, i.e. seemed a reasonable answer to the asked question. Among the reasons for non-helpfulness, we found: 30% due to misclassified intents or entities, 14% to system errors, 9% text that could be improved when replying factoid questions, 5% due to complexity of the question or not information to reply it, and 5% to specific aspects not addressed by the solution.

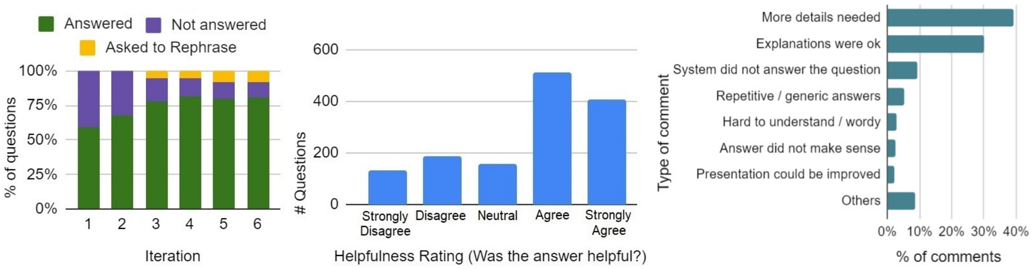


Fig. 7. Left: Distribution of replied questions, across iterations, as reported in [44]. Middle: Histogram of helpfulness rating granted by users to system answers (all iterations). Right: Types of comments by participants during corpus collection, in regard to system answers

6.4.3 Intent detection performance.

Classifiers trained on ConvEx-DS. Bert model, batch size, Adam optimizer parameters, and splitting as reported in Sec. 6.2.2. To avoid overfitting, the most represented class was downsampled (randomly), to approximate the size of the less represented class. Comparison and assessment classifiers converged after 4 epochs, detail classifier after 5.

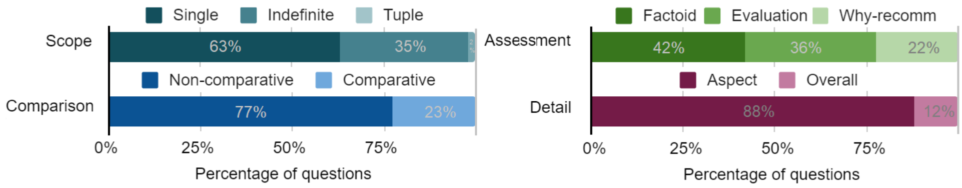


Fig. 8. Distribution of questions in ConvEx-DS (domain-related intents), as reported in [44].

Table 3. F1-scores (weighted average) of classifiers of different dimensions, trained and tested on both auxiliary datasets and ConvEx-DS, as reported in [44].

Dimension	**Dataset**	**F1**
Comparison	Jindal and Liu [51] [Training, Testing]	0.87
	Jindal and Liu [51] [Training], ConvEx-DS [Testing]	0.88
	ConvEx-DS [Training, Testing]	0.92
Assessment	Bjerva et al. [8] [Training, Testing]	0.93
	Bjerva et al. [8] [Training], ConvEx-DS (without why-recomm) [Testing]	0.60
	ConvEx-DS [Training, Testing]	0.91
Detail	WoOz augmented [Training, Testing]	0.98
	WoOz augmented [Training], ConvEx-DS [Testing]	0.90
	ConvEx-DS [Training, Testing]	0.92

Dimensions comparison, assessment and detail. To verify the performance of classifiers, we calculated the F1 score, a measure of classification accuracy. We tested accuracy in 3 different steps: 1) performance of models trained on auxiliary datasets [8, 51], used for the system used in corpus collection. 2) We tested these models using our newly obtained annotated data, ConvEx-DS. 3) We trained and tested new classifiers, based entirely on ConvEx-DS. We report F1 scores for each dimension (*comparison*, *assessment* and *detail*). We report weighted average, to take into account the contribution of each class, which in (2) is particularly unbalanced (no downsampling of the test set was done, since balanced data was pertinent only for training).

We found that the classifier trained on Bjerva et al. [8], performed particularly poor when tested with our annotated data (ConvEx-DS). Here, we detected that 32% of questions under "evaluation" class in ConvEx-DS but classified as "non-subjective" correspond to questions regarding indefinite or more than two hotels (e.g. "which hotel has the best facilities?"), 18% corresponded to adjectives like "close", "far", "expensive", and 14% to questions of the form "what is the food like?". As of factoid questions in ConvEx-DS classified as subjective, we found 33% of questions involving indefinite or more than two hotels, and 32% regarded questions of the form "does the hotel have...".

Dimension scope. Entities (hotels) addressed in sentences were detected using the NLTK library. In order to check the accuracy of the method, 2 members of the research team checked the inferred entity for the collected corpus, and found that in 5.38% of cases, the inferences were wrong. Most of these cases corresponded to cities, or facilities recognized as entities, a drawback detected in early stages of corpus collection, thus additional validations were added to the procedure, so that these cases would not occur in future iterations.

6.5 Discussion: How users communicate their explanation needs using a CA - Validation of intent model

In this section, we discuss the main findings of the WoOz pre-study, and the ConvEx-DS collection and annotation process. The content of this section was initially reported in [43, 44].

6.5.1 Types of user-generated questions. We observed, that user-generated questions collected in studies reported in [43, 44] adhered to the explanatory objective as expected, i.e. no questions regarding other processes were asked, such as hotel booking. We also observed that users actively expressed their needs for explanation, taking the lead in formulating their own questions (not expecting the system to choose what to explain) and challenging the system's attempts at argumentation when the answers provided did not satisfy their need. As expected, participants asked factoid, evaluative, and why-recommended questions. Interests expressed in factoid questions could be handled as wish conditions, and lead to changes in recommendations' appearance (e.g. highlighting those that match the desired conditions). Additionally, as expected, users not only generated standalone questions, but also follow-up questions, which confirms our expectation that an interactive QA approach would be appropriate to keep track of context and previously referred entities.

As discussed in [44], comparing our collected inquiries with the prototypical questions from the XAI question bank by [61], we found that their why-questions had a similar objective to our why-recommended: to ask for reasons why certain predictions have been provided. However, we also observed the following differences in regard to other types of questions: 1) Questions about the type of input (e.g. "what kind of data does the system learn from?") were asked only once. 2) No questions were asked about meaning of the output (e.g., "what does the system output mean"), neither on performance (e.g. "how accurate or reliable are predictions?"). We noted that how-questions asked mostly "how the opinions are" rather than asking about the overall logic of the system. 3) No "What if?" questions were asked. However, factoid questions might implicitly ask such questions (e.g. "Which hotel has a gym?" could be considered as "what if the system takes into account that 'gym' is an important feature to me?"). These differences could be explained by the context of the task to be performed, and the type of users involved (general public vs. AI experts). However, it was somehow surprising to us that only 24 out of 1806 of the collected questions referred to the system itself, its algorithm, or the type of input used for predictions. We believe that this may have been due to: 1) Users might have perceived that the recommendations matched their preferences and that they had generally positive opinions, i.e., they did not receive very strange recommendations that raised their suspicions. 2) Decisions in the chosen domain (hotels) are not as sensitive as in the medical or credit lending domains, where understanding the system logic or input influencing the prediction is critical to the acceptance of the system arguments. 3) The perspective and opinion of others might be more relevant than details about their own inferred profile, as reported by [42] for opinionated explanations in a hotel RS.

6.5.2 Intent model validation. As addressed in [44], we set out to validate how helpful the answers were to users, as an indirect measure of model validity. In this respect, we found that system answers were evaluated as predominantly helpful. We note, however, that ratings of non-helpfulness did not necessarily imply a mismatch of the detected intention. In fact, we found that almost one-third of responses rated as non-helpful fitted the question (i.e. made sense). After a review of participants' feedback on the system answers, we found that, although many users found them helpful or "ok," the main criticism was that some of the answers lacked sufficient detail, which is consistent with findings reported in [45], that perception of explanation sufficiency was greater when options to obtain further details on system claims were offered. Thus, although the intent model seems appropriate to generate an initial or first level response, a dialog system implementation must go beyond this initial response, offering options to drill down into the details, which is consistent with our proposed scheme of explanation (Sec. 3), and the dialog policy introduced in Sec. 6.6.

In regard to the annotation task, we found a substantial agreement in all the annotated dimensions, as well as a very encouraging accuracy, when classifiers were trained on the ConvEx-DS, which leads us to conclude, that under our proposed intent model and annotation guidelines, the questions could be, to a substantial extent, unequivocally classified [44]. We note, however, the challenge of addressing the dimension assessment. In this regard, we found that the main difficulty was to classify correctly questions that could be regarded as evaluation, given their subjective nature (including expressions like "how close/far"), but for which a factual-based response could be given (e.g. "100 meters from downtown"), a similar concern raised by Bjerva et al. [8]: "a subjective question may or may not be associated with a subjective answer".

6.5.3 Performance of intent classifiers. We found that intent classifiers perform better when trained on ConvEx-DS, compared to classifiers trained on the auxiliary datasets, but tested on ConvEx-DS [44]. Here, the most striking case concerns the dataset for the detection of subjective questions (SubjQA) by Bjerva et al. [8]. The above in no way suggests anything problematic in the SubjQA itself, only that in comparison to ConvEx-DS (dimension "evaluation"), the two datasets measure rather different concepts. SubjQA addresses the subjectivity of the question asked, not whether the question involves an *evaluation* that might be subjective, as in ConvEx-DS. Thus, for example, "how is the food?" is classified as non-subjective under SubjQA, since it does not contain expressions indicating subjectivity. Thus, non-subjective under SubjQA does not necessarily imply factoid. In addition, classifiers trained in SubjQA do not work well with questions that involve some sort of comparison between multiple items, since the SubjQA only involves questions addressing single items, for which an answer could be found in a single review.

6.6 Conversational Agent: Implementation

We implemented ConvEx, a dialog system for explainable RS, which follows the finite-state approach, where the dialog state has a fixed set of possible transitions to other states. The architecture of ConvEx corresponds to the prevailing dialog management architecture, according to [38]. Figure 9 illustrate the components involved in the solution, namely:

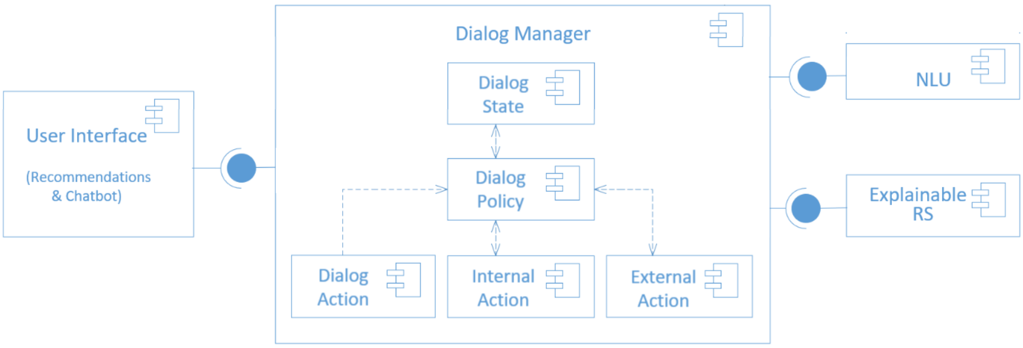


Fig. 9. ConvEx component diagram, reflecting the prevailing dialogue management architecture [38].

Dialog Manager. This component is responsible for the dialog state and flow, preserving the dialog context, and deciding on the next action to be executed by the chatbot. In turn, it involves the following components:

- Dialog State: This component allows to keep dialog context, i.e. intent and entities (hotels) addressed at each step of the dialog, for example to be able to answer follow-up questions, in case the hotel asked is not explicitly mentioned in current users' question.
- Dialog Policy: The dialog management policy defines the set of valid actions that can be executed by the chatbot given a dialog state. It is implemented in the system as a xml file, and it is depicted in figure 10. The dialog policy

can lead to the execution of different types of actions: dialog actions (to show the user a text output, e.g. the answer to a question, or to provide buttons to enable further user actions), and external actions (e.g. to invoke external REST methods, e.g. to get the list of customer comments given a hotel and a feature).

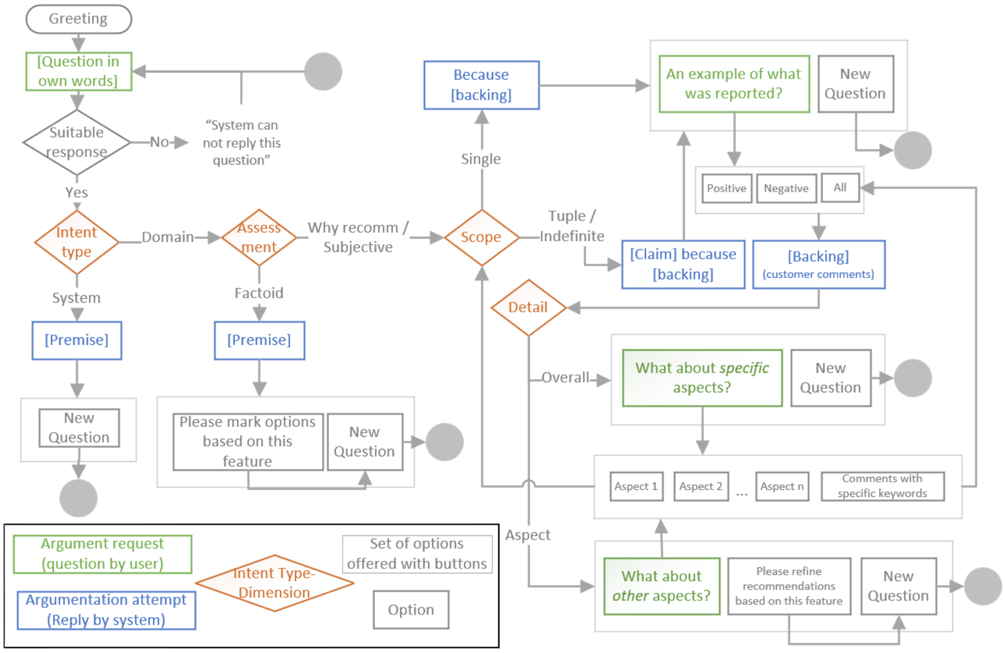
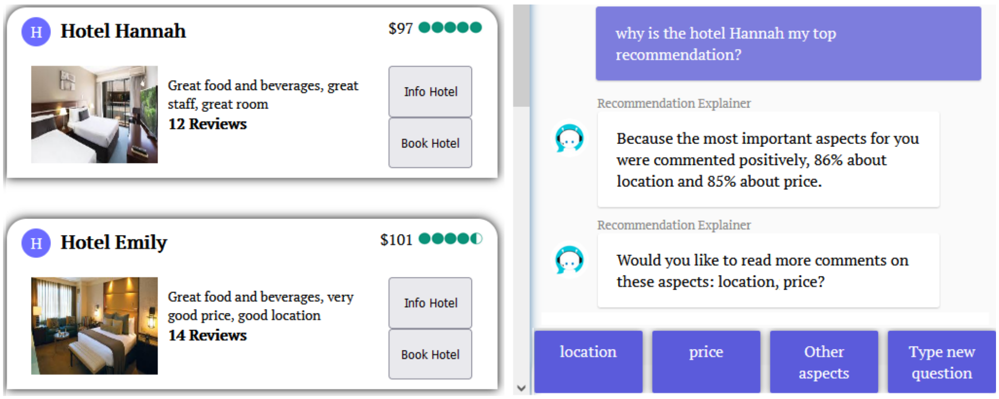

Greeting
[Question in own words]
Suitable response
No
"System can not reply this question"
Yes
Intent type
Domain
Assess ment
Why recomm / Subjective
Scope
Single
Because [backing]
An example of what was reported?
New Question
Positive
Negative
All
Tuple / Indefinite
[Claim] because [backing]
[Backing] (customer comments)
Detail
Overall
What about *specific* aspects?
New Question
Aspect 1
Aspect 2
...
Aspect n
Comments with specific keywords
Aspect
What about *other* aspects?
Please refine recommendations based on this feature
New Question
System
[Premise]
New Question
Factoid
[Premise]
Please mark options based on this feature
New Question
Argument request (question by user)
Argumentation attempt (Reply by system)
Intent Type-Dimension
Set of options offered with buttons
Option

Fig. 10. Dialog management policy for conversational explanations in review-based RS.

User interface. The user interface was implemented using the Python-based web framework Flask. It consists of two sections, as depicted in Fig. 11: list of recommendations and information about each hotel is displayed (left), and the chatbot section (right).

Rest-APIs.

- NLU: Natural language understanding module, in charge of detecting intent, and entities. This module was trained on ConvEx-DS (see Sec. 6.4.3), for which we leveraged the language model BERT [26], and procedures from the library NLTK [68] to identify the entities (particularly the tokenizer and the part of the sentence (POS) tag methods).
- ExplainableRS: We extended the implementation of the RS detailed in Sec. 4, and the method in 6.2.2, by exposing the following as REST methods: GetReply (to obtain answers to users questions), GetComments (to obtain excerpts from customers given a hotel, a feature and a polarity), GetHotelsFeature (to obtain hotels which offers a certain feature), GetReviews (to obtain reviews given a hotel) and GetRecommendations (to get recommendations, given a set of preferences chosen by the user).



Hotel Hannah $97 ●●●●●

Great food and beverages, great staff, great room

12 Reviews

Hotel Emily $101 ●●●●◐

Great food and beverages, very good price, good location

14 Reviews

why is the hotel Hannah my top recommendation?

Recommendation Explainer

Because the most important aspects for you were commented positively, 86% about location and 85% about price.

Recommendation Explainer

Would you like to read more comments on these aspects: location, price?

Fig. 11. ConvEx user interface.

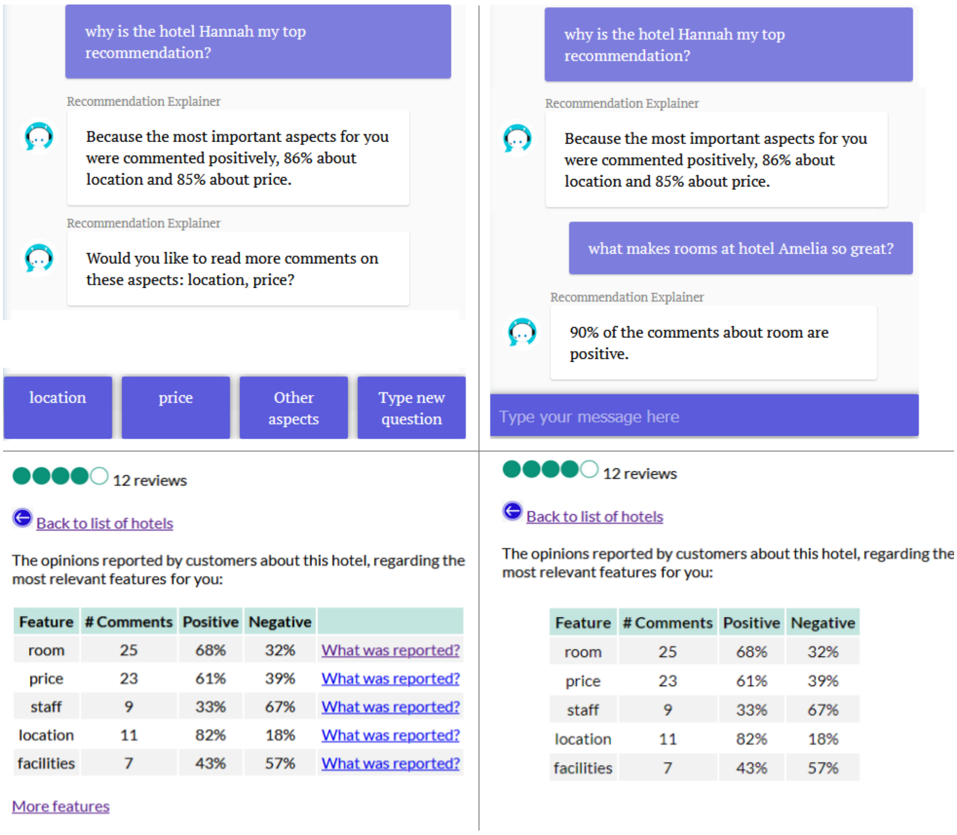
Table 4. Main differences between conditions compared in user study.

Interface Type	
GUI Navigation	**Natural language conversation**
Questions: - Fixed (“why recommended”, “what was reported on feature”), as hyperlinks or buttons *Answers:* - Table with 5 to 10 most relevant aspects for the user, # of comments and % of positive and negative comments, per aspect. - Customer comments, filtered by aspect/features and polarity.	*Questions:* User written questions, using own words. *Answers:* - Depending on detected intent (details in section 6.2.2). Answers to subjective and why-recommended questions involve % of positive and negative comments per aspect. - Customer comments, filtered by aspect/features and polarity. - Refined recommendations, for comparative and factoid questions.
Interactivity degree	
High	**Low**
- User can access arguments (customer comments), which support explanation attempts with % of positive/negative opinions per aspect. - Refined recommendations (filter hotels offering a feature), when questions are factoid (only ConvEx)	- User can not access to filtered customer comments associated to explanations. - nor filtering hotels offering a feature.

7 COMPARISON OF GUI NAVIGATION AND NATURAL LANGUAGE CONVERSATION INTERFACES

We conducted a user study to compare users’ perception of review-based RS providing explanations under different types of interfaces and interactivity degrees. Thus, we evaluated users’ perception of the system in terms of the dependent variables (DVs) transparency, effectiveness, and trust, as well as users’ assessment of the explanations in terms of explanation quality. As independent variables (IVs) we considered the factors interface *type* and interactivity *degree*. Possible values of IV *type* are: interactive *GUI navigation* and *natural language* conversation. Possible values of IV *interactivity* are: *high* and *low*. The study follows a 2x2 between-subjects design, and each participant was assigned randomly to one of four conditions (combination of *degree* and *type*, see Fig. 12). Table 4 summarizes the differences between the conditions. As covariates we considered the scores for the user characteristics decision making style (rational versus intuitive) and visualization familiarity.

Pu et al. [87] demonstrated causal relationships among RS evaluation constructs. They found that a positive perception of explanation quality by the user can lead to a positive assessment of transparency, which in turn can lead to a positive



why is the hotel Hannah my top recommendation?

Recommendation Explainer

Because the most important aspects for you were commented positively, 86% about location and 85% about price.

Recommendation Explainer

Would you like to read more comments on these aspects: location, price?

why is the hotel Hannah my top recommendation?

Recommendation Explainer

Because the most important aspects for you were commented positively, 86% about location and 85% about price.

what makes rooms at hotel Amelia so great?

Recommendation Explainer

90% of the comments about room are positive.

●●●●○ 12 reviews

Back to list of hotels

The opinions reported by customers about this hotel, regarding the most relevant features for you:

Feature	# Comments	Positive	Negative	
room	25	68%	32%	What was reported?
price	23	61%	39%	What was reported?
staff	9	33%	67%	What was reported?
location	11	82%	18%	What was reported?
facilities	7	43%	57%	What was reported?

More features

●●●●○ 12 reviews

Back to list of hotels

The opinions reported by customers about this hotel, regarding the most relevant features for you:

Feature	# Comments	Positive	Negative
room	25	68%	32%
price	23	61%	39%
staff	9	33%	67%
location	11	82%	18%
facilities	7	43%	57%

Fig. 12. The four different interfaces used in user studyManipulation of interface type in combination with interactivity degree, in user study. Top: natural language, bottom GUI navigation. Left: interactivity degree high, right: low.

assessment of trust. In addition, they found that a positive perception of information sufficiency (a factor considered in our evaluation of explanation quality) can lead to a positive perception of perceived usefulness, which is related to the definition of system effectiveness by Knijnenburg et al. [55]. Consequently, we expect a positive correlation of our evaluated variables.

Furthermore, proponents of dialog models of explanation argue that providing explanations through a conversation between the user and the system, in which the user can contest the system's arguments, can facilitate user understanding of system decisions [46, 103]. This understanding, in turn, is the basis on which the concept of transparency has traditionally been evaluated in RS (e.g. [24, 30, 87]).

Since the natural language interface allows the user to formulate a wider range of questions, even in their own words, compared to the GUI navigation approach, where questions are limited to a much smaller number of fixed options represented in links and buttons, we hypothesized that:

H1: Users' evaluation of the system is more positive when they are provided with explanations through a natural language interface, than when provided through a GUI navigation.

According to our proposal, explanations with a higher degree of interactivity allow users to contest the arguments of the system in the form of aggregated opinions, allowing to review in detail customer comments that compose such aggregation, which could facilitate the understanding of the recommendations. Thus, we hypothesized that:

H2: Users' evaluation of the system and its explanations is more positive when explanations are provided with a higher degree of interactivity.

Additionally, we hypothesized (in line with [55, 64]) that a number of user characteristics may moderate the effect of interactive functionalities on the evaluation of the system and its explanations. Particularly, users with greater visualization familiarity have a greater ability and habit of processing aggregated information through graphical formats, such as those included in the GUI navigation condition, and users with lower visual abilities might benefit less from a presentation based on tables or graphics [53, 92]). Thus, we hypothesized that:

H3: Users more familiar with data visualization will evaluate GUI navigation explanations more positively.

Finally, since users with a predominant rational decision-making style have a preference for evaluating information extensively during decision making [37], as reported by [45], and facilitated by our argumentative explanation design, we hypothesized that:

H4: Users with a predominantly rational decision style will evaluate explanations under our interactive approach more positively than less rational decision makers.

7.1 User study

7.1.1 Questionnaires.

Evaluation. We utilized items from [55] to evaluate system effectiveness (internal reliability Cronbach's $\alpha = 0.78$, construct *perceived system effectiveness*, system is useful and helps the user to make better choices). We used items from [74] to evaluate trust in the system ($\alpha = 0.80$, sub-constructs *trusting beliefs*, user considers the system to be honest and trusts its recommendations; and *trusting intentions*, user willing to share information).

To measure users' evaluation of the RS transparency, we used own items ($\alpha = 0.86$), involving the sub-constructs: input ("what kind of data does the system learn from"), output ("what kind of output does the system give"), functionality ("how / why does the system make predictions") and interaction (what if / how to be that, "what would the system predict if this instance changes to..."). These constructs are related to the main components of RS structure [50, 70], and to the categories of typical user questions related to AI algorithms [61]. Items and factor loadings are reported in Appendix B.

We used the user experience items (UXP) of [56] to address reception of explanations, which we will refer to as *explanation quality* (internal consistency of the construct: $\alpha = 0.88$), comprising: confidence (explanations makes user confident that she/he would like the recommended item), transparency (explanation makes the recommendation process clear), satisfaction (user would enjoy a system if explanations are presented this way, -item adapted from original-), and persuasiveness (explanations are convincing). We added an item adapted from [28] (explanations provided are sufficient to make a decision) to evaluate explanation sufficiency.

All items were measured on a 1-5 Likert-scale (1: Strongly disagree, 5: Strongly agree). In addition, the following open-ended questions were asked: "Please provide the reasons why you chose the hotel you did. A bonus of up to £0.3 will be paid depending on the quality of this response.", "Please let us know about your overall opinion about the explanations provided or how they could be improved", and "How would you explain to a friend, in your own words, how the system generates recommendations?". Questionnaire items are listed in Appendix, section B.

User characteristics. We used all the items of the decision style scale proposed by [37], which is a well-validated instrument for rational and intuitive decision making. We used the visualization familiarity items proposed by [56] (α = 0.87). All items were measured with a 1-5 Likert-scale (1: Strongly disagree, 5: Strongly agree). Questionnaire items are listed in Appendix, section B.

7.1.2 Participants. We recruited 162 participants (82 female, mean age 29.16 and range between 18 and 66) through the crowdsourcing platform Prolific. We restricted the task to workers in the U.S and the U.K., with an approval rate greater than 98%. Participants were rewarded with £1.7 plus a bonus up to £0.30 depending on the quality of their response to the question about reasons why they chose a certain hotel, set at the end of the survey, aiming to achieve a more motivated hotel choice by participants, and encouraging effective interaction with the system. Time devoted to the task (in minutes) was M=16.39, SD=5.06 (time of interaction with the system was M=5.98, SD= 3.48).

We applied a quality check to select participants with quality survey responses. We included attention checks in the survey, e.g. "This is an attention check. Please click here the option 'Disagree'". The responses of 28 of the 190 initial participants were discarded, on the basis of failing validation checks, suspicious response patterns (e.g. the same score for all questions on the same page) or a very low time using RS (less than 40 seconds), for a final sample of 162 subjects, statistical power of 0.91, α =0.05 (power values above conventional for adequacy of .80 are considered acceptable [86]; 'a priori' analysis was performed in G*power [32]).

7.1.3 Procedure. First, the participants answered the questionnaires on user characteristics. They were asked to report their 5 most important aspects when looking for a hotel, sorted by importance. They were presented with instructions on how to complete the task: 1. A list of 10 recommended hotels would be presented (as a result of a hypothetical hotel search already performed using a RS, i.e., no search filters were offered). 2. They had to evaluate the different options presented, come to a decision on which hotel appeals to them the most, and back in the survey, they had to indicate the reasons for the chosen option.

They were also presented with specific instructions and screenshots on how to interact with the system and obtain explanatory information, depending on the condition assigned (these were also included within the system as a reminder). In the conditions with a natural language interface, a list of 9 examples of the type of questions they could ask to the chatbot was provided. Participants were then presented with a cover story, to establish a common starting point in terms of travel motivation (a holiday trip). The participants then used the application and filled out the system evaluation questionnaire.

7.2 Data analysis

We assessed the effect that IVs (interactivity degree and interface type) may have on the evaluation of the system and its explanations (DVs explanation quality, and system transparency, effectiveness and trust), and to what extent the covariates (user characteristics: rational and the intuitive decision making styles and visualization familiarity) could influence such evaluation.

Evaluation scores (DVs' scores) and covariates scores for each individual were calculated as the average of the reported values for the scale items.

Given that DVs are continuous and correlated (see Table 5), a MANCOVA analysis was performed. Subsequent ANCOVAs were performed to test main effects of IVs and covariates, as well as the effect of interactions between them. Q-Q plots of residuals were checked to validate the adequacy of the analysis. Regression analysis was used to evaluate moderation effects between covariates and IVs on DVs. An additional MANCOVA analysis was performed to evaluate

Table 5. Mean values and standard deviations of evaluation of RS, per interface type and interactivity degree (n=162), p<0.05*, p<0.01**; values reported with a 5-Likert scale; higher mean values correspond to a positive evaluation of the RS. Pearson correlation matrix, p<0.001 for all coefficients.

Variable	*Interface type*				*Interactivity degree*				*Correlation*			
	Natural language		GUI Navigation		Low		High		Variable			
	M	SD	M	SD	M	SD	M	SD	1	2	3	4
1. Expl. Quality	3.75	0.71	3.82	0.68	3.67	0.76	3.90*	0.60	-			
2. Transparency	3.62	0.54	3.63	0.52	3.50	0.59	3.74**	0.45	0.52	-		
3. Effectiveness	3.91	0.69	3.88	0.62	3.81	0.70	3.98	0.61	0.64	0.53	-	
4. Trust	3.87	0.57	3.79	0.47	3.75	0.55	3.91*	0.48	0.70	0.62	0.81	-

the effect of the IVs on sub-constructs of the transparency scale: input, output, functionality and interaction. Q-Q plots of residuals were also checked to validate the adequacy of the analysis.

7.3 Results

7.3.1 *Quantitative evaluation.*

Evaluation of RS and User Characteristics Scores. The average evaluation scores by interface type and interactivity degree are shown in Table 5. Distributions of the scores of rational (*M* = 4.26, *SD*= 0.60) and intuitive (*M* = 2.87, *SD*= 0.73) decision making styles, and visualization familiarity (*M* = 2.90, *SD*= 1.00) are depicted in Fig. 13.

Interactivity degree. We found a significant multivariate effect of interactivity degree on system evaluation F(4,152) = 2.51, p = .044. Univariate tests revealed that degree of interactivity significantly influences the evaluation of explanation quality F(1,155) = 4.53, p = .035, transparency F(1,155) = 9.34, p = .003, and trust F(1,155) = 3.97, p = .048. In all these cases, the average of every variable was higher for the high condition than for low condition.

Interface type. We found no significant main effect of interface type, nor of interaction between interface type and interactivity degree. However, we found that user traits might mediate the effect of type on system evaluation, as explained below.

Visualization familiarity. We found a significant multivariate effect of this variable, F(4,152) = 2.47, p = .047. Univariate tests revealed a main effect of this variable on explanation quality F(1,155) = 5.65, p = .019, and on transparency F(1,155) = 5.95, p = .016. Here, a positive trend was observed between these variables and visualization familiarity, i.e. the higher the visualization familiarity score, the higher the perceived explanation quality, and the transparency.

Moderation between interface type and visualization familiarity. These variables accounted for a significant amount of variance in explanation quality, R2 = .068, F(3,158) = 3.83, p = .011. Examination of the interaction plot (Fig. 13b) showed an enhancing effect that as visualization familiarity increased, positive perception of explanation quality under the condition GUI navigation increased, while under the natural language condition, explanation quality remained rather constant for different values of visualization familiarity.

Rational decision-making. We found a significant multivariate effect of this variable, F(4,152) = 4.17, p = .003. Univariate tests revealed a main effect of this variable on explanation quality F(1,155) = 5.82, p = .017, and transparency F(1,155) =

13.27, p < .001. Here, a positive trend was observed between these variables, i.e. the higher the rational decision-making, the higher the perceived explanation quality, and the transparency.

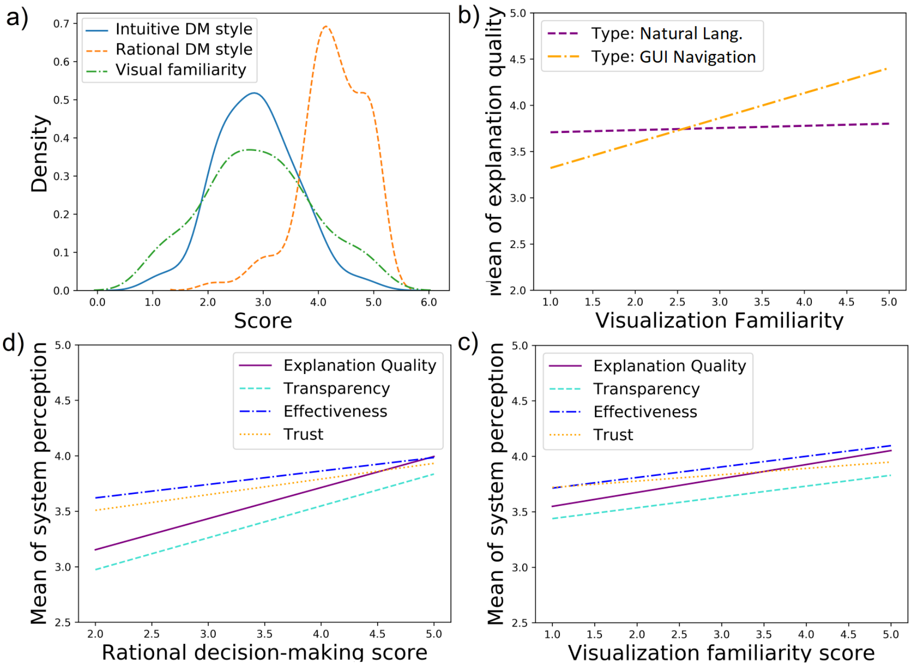


Fig. 13. a) Kernel density estimate of user characteristics scores: rational and intuitive decision making styles, and visualization familiarity. b) mediation effect between visualization familiarity and interface type on explanation quality (fitted means of individual scores). c) Effect of rational decision-making style on the evaluation of the system (fitted means of individual scores). d) Effect of visualization familiarity on the evaluation of the system (fitted means of individual scores).

7.3.2 Quantitative evaluation (sub-constructs transparency). The average evaluation scores for every sub-construct of the transparency scale are shown in Table 6.

Interactivity degree. We found a significant multivariate effect of interactivity degree on transparency when sub-constructs are addressed: F(4,152) =3.10, p = .017. Univariate tests revealed that degree of interactivity significantly influences the evaluation of input F(1,155) = 5.40, p = .021, output F(1,155) = 4.43, p = .037, functionality F(1,155) = 4.94, p = .028, and interaction F(1,155) = 9.97, p = .002. In all these cases, the average of every variable was higher for the high condition than for low condition.

7.3.3 Qualitative evaluation.

Comments and suggestions by participants. We analysed the answers of participants to the open ended question, asking for opinions about the system and its explanations, and categorized the responses of participants (Fig. 14).

Table 6. Mean values and standard deviations of evaluation of transparency sub-constructs, per interface type and interactivity degree (n=162), p<0.05*, p<0.01**; values reported with a 5-Likert scale; higher mean values correspond to a positive evaluation of transparency. Pearson correlation matrix, p<0.001 for all coefficients.

Variable	*Interface type*				*Interactivity degree*				*Correlation*			
	Natural language		GUI Navigation		Low		High		Variable			
	M	SD	M	SD	M	SD	M	SD	1	2	3	4
1. Input	3.67	0.67	3.70	0.66	3.57	0.67	3.80*	0.64	-			
2. Output	3.77	0.68	3.84	0.65	3.70	0.73	3.91*	0.57	0.59	-		
3. Functionality	3.61	0.61	3.65	0.58	3.53	0.68	3.73*	0.48	0.66	0.60	-	
4. Interaction	3.41	0.83	3.22	0.84	3.12	0.79	3.51**	0.85	0.34	0.31	0.37	-

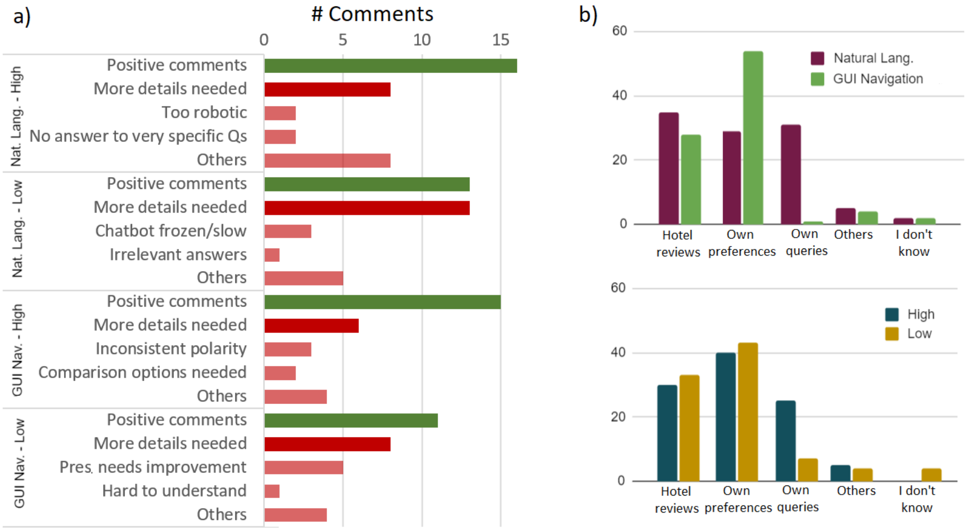


Fig. 14. a) Type of comments and suggestions by participants, about the system and its explanations, per condition. Positive comments in green, negative comments in red (light red for less frequent comments). b) Type of information input used by the system to generate recommendations, according to participants, top : by interface type, bottom: by interaction degree

Type of information input used for recommendations. We analysed the answers of participants to the open ended question, asking for how they would explain to a friend how the system generated the recommendations. We categorized the responses of participants, according to the type of information related in participants' comments (Fig. 14).

7.4 Discussion: Comparison of GUI navigation and natural language conversation interfaces

7.4.1 Quantitative evaluation. Providing interactive explanations in RS under our proposed approach, and using both types of interface (natural language and GUI navigation), proved to improve users' assessment of the system, given the predominant positive evaluation of the two tested systems and their explanations by the participants. However, we found no significant difference in the evaluation of the two types of interfaces, unless user characteristics, such as visualization familiarity are taken into account. Particularly, we could not confirm our **H1** (users' evaluation of the system is more positive when they interact with an explanatory natural language interface).

The natural language interface may entail advantages for certain users. Under this approach, however, the user must take the initiative to formulate their own questions, which competes with a more proactive position on the part of the system in the case of GUI navigation, where it provides a more comprehensive summary of opinions (covering a greater number of aspects at the same time in a table), without expecting the user to inquire about specific aspects. Additionally, visual arguments (a combination of visual and verbal information, e.g. using a table as in the GUI navigation) may have a greater "rhetorical power potential" than verbal arguments, due (among others) to their greater immediacy (possibility of quick processing) [9].

The above is further reinforced by our finding in relation to user characteristics, that users with greater visualization familiarity reported a more positive evaluation of explanation quality, when explanations were provided using GUI navigation, thus confirming our **H3**. Here, the main design feature that seems to play a significant role in this difference, is the display of a global evaluation of customer opinions, referencing a wider range of aspects (up to 10) in a single table. Such display contrasts to the natural language approach, in which aggregations of opinions were indicated only up to two aspects (following the principle of brevity suggested by [79], answers should be provided as concisely as possible). Also, such type of answers were only provided under the intent types assessment *subjective* and *why-recommended*, and detail *overall* (no aspect in particular is asked), this latter only representing the 12% of the users' utterances (Fig. 8). Thus, the natural language approach focuses on answering specific user questions, but at the same time involves a shortcoming: it makes more difficult to users to have a look at aspects that may not be of their greatest interest, but may be decisive in making a final decision, especially if they have customer reports that were not so favorable.

Using a graphical user interface can result in shorter times to access the required information. Additionally, explaining recommendations by means of GUI navigation implies an advantage over the natural language interface, since, as in the case of direct-manipulation style interfaces [33], they can provide "consistent and concrete representations of data, operations and system states" [11], as well as clear options suggesting what the user can do next. On the other hand, using natural language interfaces might increase the users' expectations in regard to what a CA can actually perform [48]. The above can, however, be mitigated by a continuing use of the application, as well as presenting new users with examples on how to interact with the CA, as well as the type of questions the user can ask [48], guidelines we followed in our user study. The above contributed, in our opinion, to the predominantly positive evaluation we observed in the natural language condition.

Our results show that higher degree of interactivity has a significantly positive effect on users' evaluation of the system, independent of the type of interface, particularly in terms of system transparency and trust, as well as of explanation quality, compared to explanations with a lower degree of interactivity, thus confirming our **H2**. The above in turn confirms the suitability of our proposal for explanations as interactive argumentation, inspired by dialog models of explanation, which enables users to question initial explanation attempts provided by the system, in particular the aggregate representation of positive and negative customer opinions. Here, active control and two-way as addressed in our proposal (in both natural language and GUI navigation interfaces) seem to play a role in the observed effect, namely: active control, as users can exert control over the argumentative content to be displayed by the system; two-way communication, as users can inform the system which statements require further backing, and which features are of momentary relevance to them, thus contributing to a better understanding of explanations, as sustained by dialog models of explanation [46, 103]. We note, however, that despite the significant multivariate effect of interactivity degree, and the higher mean reported for effectiveness under the high degree, this observed effect was not significant. We believe this might be due to, despite providing more information under the high degree condition, and in line with some participants' comments, more detailed information was still needed to make a final decision, especially concerning very specific factoid questions.

The use and perceived benefit of interaction options might be influenced by individual differences, as discussed by [64]. We found that the way people process information when making decisions play a significant role in the evaluation of explanations as natural language conversation, as also reported in [45] for GUI navigation. In particular, we found that a more rational decision-making style influences significantly the evaluation of the system and its explanations. Here, we observed that participants who reported higher rational decision making scores reported significantly higher scores on transparency and explanation quality, independent of the interface type or interaction degree, thus confirming our **H4**. The above could be explained by the propensity of people with a predominant rational decision-making style, to search for information and evaluate alternatives exhaustively [37], which is facilitated by our approach. This suggests that interactive review-based explanations seem to benefit more the users who tend to evaluate information thoroughly when making decisions, compared to users who use a more shallow information-seeking process.

7.4.2 Qualitative evaluation. When analyzing the comments reported by users regarding their perception and suggestions for improving the system, we found, in line with the quantitative evaluation, that more positive comments were reported for the conditions under the high degree of interactivity than for the low degree. Here, most of the negative comments reported by participants refer to the lack of additional details: 1) under the low interactivity condition, a lack of access to customer comments supporting the explanations, and 2) under the natural language interface condition, the lack of specific details in answers to factoid questions (such as, for example, type of menu offered in hotel's restaurant, or distance to the beach).

In relation to the open-ended question, in which participants were asked to indicate how they would explain to a someone else how the system generates recommendations, we noted an overall fair understanding of what kind of information the system uses to generate recommendations, that is, the participants' responses mostly reflected the input used, rather than specific details about the algorithm, for which our explanations do not provide details. On one hand, the number of responses mentioning that the system uses hotel reviews and ratings is very similar across the conditions. On the other hand, a salient difference can be observed in comments referring to preferences as a base for recommendations, in particular between different interface types. Here, participants under the GUI navigation condition reported much more comments indicating that recommendations were based on their own preferences.

The above seems to be the effect of showing explanations using a list view of preferred aspects and aggregated customer opinions, within the same table, while in the natural language case, the responses indicating that recommendations are based on preferences are only provided for questions that do not address an aspect in particular (around 12% of asked questions), e.g. Q: "Why does Hotel Amelia have the highest rating?" A: "Because the most important aspects for you were commented positively, 90% about room and 69% about location". However, the lack of awareness of preferences used as input of recommendations might be mitigated by users being aware that recommendations are generated based on their queries. That is, in the natural language condition, the understanding that recommendations can be refined as a result of the conversation seems fostered. The above seems even more pronounced under the natural language condition with a high degree of interactivity, where in addition to a simple exchange of textual questions and answers, additional options are offered, for example, to refine the recommendations given a certain aspect or feature. This qualitative finding is also consistent with the significant difference observed in the scores reported for the *interaction* sub-construct of the transparency scale, i.e., the user can better identify what actions to take and what to change to obtain better recommendations, in the case of explanations with a high degree of interactivity.

Finally, only a low number of comments indicated that they did not know how the system generated the recommendations (4 comments, all under low degree of interactivity), or others (9 comments pointing to general hotel info, or its price), or indicated reasons that had nothing to do with the actual functioning of the system (only 1 comment related to system using preferences information from 3rd parties), which suggests that our interactive approach to explanations contributes to a great extent to the understanding of the information that the system uses to generate its recommendations.

8 CONCLUSIONS AND OUTLOOK

We discussed in Sec. 3 a scheme for explanations as interactive argumentation in review-based RS, inspired by dialog explanation models and formal argument schemes, that allows users to go from aggregated customer opinions to detailed extracts of individual reviews, in order to facilitate a better understanding of the claims made by the RS. We found that providing interactive explanations in RS based on the proposed scheme proved to be an effective means for improving users' evaluation of the system, both in the GUI navigation and natural language conversation variants of the RS implemented. We found, however, that users with greater visualization familiarity reported a more positive perception of explanation quality, when explanations were provided through GUI navigation.

We also found that providing a higher degree of interactivity in explanations contributed to a more positive evaluation of transparency and trust in the system, independent of the interactive interface type. We furthermore found that individual differences in terms of user characteristics (e.g. decision-making style, and visualization familiarity) may lead to differences in the evaluation of the proposed implementation.

Based on our findings concerning language-based conversational explanations, we conclude that our proposal of a dimension-based intention model is a valid approach to represent user queries in the context of explanatory RS. By providing ConvEx-DS, a data set annotated based on this model, we also hope to contribute to the development of future dialog systems that support conversational explanations in RS.

8.1 Practical implications

Our findings lead to the following practical implications:

- Providing explanations resembling a conversation between user and system and user can contribute to a better evaluation of RS. This conversation can be enabled either through traditional web browsing options, or through a natural language conversation, allowing the user to indicate their explanation needs using their own words, using for example, a conversational agent.
- Providing both a global view of opinions and actual customer comments, which can be filtered by aspects or specific features at will, can positively impact the perception of the transparency of the system, while avoiding overwhelming users with irrelevant information in a single step.
- We suggest providing options for the user to choose between their preferred visualization and navigation style, i.e. GUI navigation or a chatbot-style conversation, rather than simply offering one or the other.

8.2 Limitations and future work

Despite our motivating results, it is important to note the limitations imposed by the discussed proposal. First, our approach involves providing excerpts of customer reviews as backing for aggregation-type explanations, filtered by aspect and sentiment. However, we did not implement methods for sorting customer comments according to their helpfulness or argumentative quality. Thus, we plan in the future to integrate techniques that will allow us to take the quality of comments into account, as well as to evaluate their effect, both on the accuracy of recommendation prediction and on the evaluation of the system by end-users.

Furthermore, our proposal has been specifically applied to the hotel domain. However, we suppose that our approach will generalize properly to domains related to experience goods [82], where the search for information during decision making relies heavily on word-of-mouth [54, 83], as is the case, for example, for restaurants. Thus, in future work we will evaluate the use of our explanation scheme in domains other than hotel, as well as the validity of our intention model, and the usefulness of ConvEx-DS.

In regard to our conversational approach. addressing intent detection as a text classification problem by means of an intent classification model, allows to provide answers that approximate the information need expressed by the user. However, the approach is insufficient when dealing with questions that are too specific, particularly in regard to factoid questions. Consequently, the development of a conversational agent with explanatory purposes in RS should not only rely on the underlying RS algorithm, customer reviews or hotels metadata (as in our developed system), but should also integrate further sources of information, e.g. external location services, in order to provide specific details, such as surroundings, distances to places of interest or transport means, in case these are not found in customer reviews or metadata.

Furthermore, as reported in Sec. 7, participants under the natural language condition acknowledged that the system refined the recommendations as a result of the conversation with the system. However, the system was not designed as a critiquing RS as such, where users can directly express their preferences by modifying features of recommended items. In particular, the dialog policy designed as the basis of the ConvEx system focused on the support of the explanation process, rather than on the elicitation of preferences. Therefore, we plan in the future to evaluate and implement a system that blends the functionality of critiquing systems with that of conversational explanations to improve user experience.

Finally, conversations in ConvEx are based on a dialog manager that falls into the handcrafted category of dialog management approaches, where the state and policy are defined as a set of rules defined by developers and domain experts. While suitable for our proposed dialog policy, and for testing our hypothesis in Sec. 7, more flexible approaches would be needed to extend the policy for supporting preference elicitation as well as a broader range of user goals such as item search, customer service or booking. End-to-end neural network approaches seem promising for this purpose, as they allow for a more dynamic modeling of possible dialogues with different goals (as proposed by [65]), an alternative that we consider exploring in the future.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18*. 1–18.

[2] Abdallah Arioua and Madalina Croitoru. 2015. Formalizing Explanatory Dialogues. *Scalable Uncertainty Management* (2015), 282–297.

[3] Abdallah Arioua, Madalina Croitoru, Laura Papaleo, Nathalie Pernelle, and Swan Rocher. 2016. On the Explanation of SameAs Statements Using Argumentation. *Scalable Uncertainty Management* (2016), 51–66. https://doi.org/10.1007/978-3-319-45856-4_4

[4] Roland Bader, Wolfgang Woerndl, Andreas Karitnig, and Gerhard Leitner. 2012. Designing an explanation interface for proactive recommendations in automotive scenarios. In *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization (UMAP'11)*. 92–104.

[5] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 717–725.

[6] Jamal Bentahar, Bernard Moulin, and Micheline Belanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review* 33, 3 (2010), 211–259.

[7] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend?: User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces.* 287–300.

[8] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP.* 5480–5494. https://doi.org/10.18653/v1/2020.emnlp-main.442

[9] J. Anthony Blair. 2012. The Possibility and Actuality of Visual Arguments. *Tindale C. (eds) Groundwork in the Theory of Argumentation* 21 (2012), 205–223.

[10] Marco De Boni and Suresh Manandhar. 2020. Implementing clarification dialogues in open domain question answering. *Natural Language Engineering* 11, 4 (2020), 343–361. https://doi.org/10.1017/S1351324905003682

[11] Susan E Brennan. 1998. The grounding problem in conversations with and through computers. In *Social and cognitive psychological approaches to interpersonal communication*, S. Fussell and R. Kreuz (eds.) (Eds.). Hillsdale, NJ: Lawrence Erlbaum, Chapter 9, 210–225.

[12] Andrei Broder. 2002. A taxonomy of web search. *ACM SIGIR Forum* 36, 2 (2002), 3–10.

[13] Dimitrios Buhalis and Emily Siaw Yen Cheng. 2020. Exploring the Use of Chatbots in Hotels: Technology Providers Perspective. *Information and Communication Technologies in Tourism* (2020), 231–242. https://doi.org/10.1007/978-3-030-36737-4_19

[14] Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2013. Multi document summarization of evaluative text. In *Computational Intelligence*, Vol. 29. 545–574.

[15] Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. In *Artif. Intell.*, Vol. 170. 925–952.

[16] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient Intent Detection with Dual Sentence Encoders. In *arXiv:2003.04807.*

[17] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee.* 1583–1592.

[18] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35. https://doi.org/10.1145/3166054.3166058

[19] Li Chen and Pearl Pu. 2014. Critiquing-based recommenders: survey and emerging trends. 22, 1–2 (2014), 3085–3094.

[20] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–12.

[21] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 16.* 815–824. https://doi.org/10.1145/2939672.2939746

[22] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic Generation of Natural Language Explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion.* 57:1–57:2.

[23] Alice Coucke, Alaa Saade, Adrien Ball, Theodore Bluche, Alexandre Caulier, David Leroy, and Clement Doumouro. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. In *ArXiv, abs/1805.10190.*

[24] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User-Adap. Inter.* 18 (2008), 455–496.

[25] Nils Dahlback, Arne Jonsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st Int. Conference on Intelligent User Interface.* 193–200.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019).

[27] Ruihai Dong, Michael P. O Mahony, and Barry Smyth. 2014. Further Experiments in Opinionated Product Recommendation. In *Case Based Reasoning Research and Development.* Springer International Publishing, 110–124.

[28] Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2020. Explaining recommendations by means of aspect-based transparent memories. In *Proceedings of the 25th International Conference on Intelligent User Interfaces.* 166–176.

[29] Tim Donkers and Jürgen Ziegler. 2020. Leveraging Arguments in User Reviews for Generating and Explaining Recommendations. *Datenbank-Spektrum* 20, 2 (2020), 181–187.

[30] Simon Dooms, Toon De Pessemier, and Luc Martens. 2011. A user-centric evaluation of recommender algorithms for an event recommendation system. In *Proceedings of the RecSys 2011: Workshop on Human Decision Making in Recommender Systems (Decisions RecSys 11) and User-Centric Evaluation of Recommender Systems and Their Interfaces - 2 (UCERSTI 2) affiliated with the 5th ACM Conference on Recommender Systems (RecSys 2011).* 67–73.

[31] Michael J. Driver, Kenneth E. Brousseau, and Phil L. Hunsaker. 1990. The dynamic decision maker. (1990).

[32] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39 (2007), 175–191.

[33] Claes Fornell and David F. Larcker. 1982. The future of interactive systems and the emergence of direct manipulation. *Behavior and Information Technology* 1 (1982), 237–256.

[34] Gerhard Friedrich and Markus Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. 32, 3 (2011), 90–98.

[35] Ralph Grishman and Beth Sundheim. 2020. Message understanding conference- 6: A brief history. In *Proceedings of the 28th International Conference on Computational Linguistics*. 466–471.

[36] Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. In *Computational Linguistics 43*, Vol. 1. 125–179.

[37] Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The Development and Validation of the Rational and Intuitive Decision Styles Scale. *Journal of Personality Assessment* 98, 5 (2016), 523–535.

[38] Jan-Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2019. Approaches for Dialog Management in Conversational Agents. *IEEE Internet Computing* 23, 2 (2019), 13–22. https://doi.org/10.1109/MIC.2018.2881519

[39] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *In Proceedings of the workshop on Speech and Natural Language - HLT '90*. 96–101.

[40] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[41] Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*. https://doi.org/10.1145/3386392.3399302

[42] Diana C Hernandez-Bocanegra and Juergen Ziegler. 2020. Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics. *Journal of Interactive Media* 19, 3 (2020), 181–200. https://doi.org/10.1515/icom-2020-0021

[43] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *3rd Conference on Conversational User Interfaces (CUI '21)*. 1–11. https://doi.org/10.1145/3469595.3469596

[44] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. ConvEx-DS: A dataset for conversational explanations in recommender systems. In *Proceedings of IntRS 21: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 1–18. http://ceur-ws.org/Vol-2948/paper1.pdf

[45] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Effects of Interactivity and Presentation on Review-Based Explanations for Recommendations. In *Human-Computer Interaction – INTERACT 2021*. Springer International Publishing, 597–618. https://doi.org/10.1007/978-3-030-85616-8_35

[46] Denis J. Hilton. 1990. Conversational processes and causal explanation. 107, 1 (1990), 65–81.

[47] Jian Hu, Gang Wang, Fred Lochovskyand Jian tao Sun, and Zheng Chen. 2009. Understanding user's query intent with wikipedia. In *Proceedings of the 18th international conference on World wide web - WWW '09*.

[48] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference DIS*. 895–906. https://doi.org/10.1145/3196709.3196735

[49] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54, 5 (2021), 1–36.

[50] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2011. *Recommender Systems. An introduction*. Cambridge University Press.

[51] Nitin Jindal and Bing Liu. 2006. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 06*. 244–251. https://doi.org/10.1145/1148170.1148215

[52] John F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Information Applications. In *Transactions on Office Information Systems*, Vol. 2. 26–41.

[53] John R Kirby, Phillip J Moore, and Neville J Schofield. 1988. Verbal and visual learning styles. *Contemporary Educational Psychology* 12, 2 (1988), 169–184.

[54] Lisa Klein. 1998. Evaluating the Potential of InteractiveMedia through a New Lens: Search versus Experience Goods. In *Journal of Business Research*, Vol. 41. 195–203.

[55] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. In *User Modeling and User-Adapted Interaction*. 441–504.

[56] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI 19)*. ACM, 379–390.

[57] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5686–5697.

[58] Béatrice Lamche, Ugur Adigüzel, and Wolfgang Wörndl. 2012. Interactive explanations in mobile shopping recommender systems. In *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE'14), held in conjunction with the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP'14)*. 92–104.

[59] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. Klagenfurt, Germany: SSOAR.

[60] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *32nd Conference on Neural Information Processing Systems, NeurIPS 2018*. 9725–9735.

[61] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 9042 (2020), 1–15. https://doi.org/10.1145/3313831.3376590

[62] Nathalie Rose Lim, Patrick Saint-Dizier, , and Rachel Roxas. 2009. Some challenges in the design of comparative and evaluative question answering systems. In *In Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions - KRAQ 09*. 15–18. https://doi.org/10.3115/1697288.1697292

[63] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement* 27 (1990), 247–266.

[64] Yuping Liu and L J Shrum. 2002. What Is Interactivity and Is It Always Such a Good Thing? Implications of Definition, Person, and Situation for the Influence of Interactivity on Advertising Effectiveness. *Journal of Advertising* 31, 4 (2002), 53–64.

[65] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che1y, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).

[66] Benedikt Loepp, Katja Herrmanny, and Juergen Ziegler. 2015. Blended Recommending: Integrating Interactive Information Filtering and Algorithmic Recommender Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15*. 975–984.

[67] Benedikt Loepp, Tim Hussein, and Juergen Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI 14*. 3085–3094.

[68] Edward Loper and Steven Bird. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. (2009).

[69] Samuel Louvan and Bernardo Magnini. 2020. Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*. 480–496.

[70] Jie Lu, Qian Zhang, and Guang quan Zhang. 2021. *Recommender Systems. Advanced Developments.* World Scientific Publishing.

[71] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019*. 1–9.

[72] Bella Martin and Bruce Hanington. 2012. *Universal Methods of Design.* Rockport Publishers, Beverly, MA.

[73] Philipp Mayring. 2014. Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution. (2014). Klagenfurt, Germany: SSOAR.

[74] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. In *Information Systems Research*, Vol. 13.

[75] Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human Annotated Dialogues Dataset for Natural Conversational Agents. *Appl. Sci* 10, 762 (2020), 1–16.

[76] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* (2018).

[77] Amit Mishra and Sanjay Kumar Jain. 2015. An Approach for Sentiment analysis of Complex Comparative Opinion Why Type Questions Asked on Product Review Sites. *Computational Linguistics and Intelligent Text Processing Springer LNCS* 9042 (2015), 257–271.

[78] Christof Monz. 2003. Document Retrieval in the Context of Question Answering. In *Proceedings of the 25th European conference on IR research*. 571–579.

[79] Robert J. Moore and Raphael Arar. 2018. Conversational UX Design: An Introduction. *Studies in Conversational UX Design* (2018), 1–16. https://doi.org/10.1007/978-3-319-95579-7_1 Springer International Publishing.

[80] Emanuela Moreale and Maria Vargas-Vera. 2004. A Question-Answering System Using Argumentation. *MICAI 2004: Advances in Artificial Intelligence* (2004), 400–409. https://doi.org/10.1007/978-3-540-24694-7_41

[81] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Intelligent User Interfaces (IUI 16)*, Vol. 2. 256–260.

[82] Phillip Nelson. 1970. Information and Consumer Behavior. 78, 2 (1970), 311–329.

[83] Philip J. Nelson. 1981. Consumer Information and Advertising. In *Economics of Information*. 42–77.

[84] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model User-Adap* 27 (2017), 393–444.

[85] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 135–143. https://doi.org/10.1145/3349537.3351899

[86] Marco Perugini, Marcello Gallucci, and Giulio Costantini. 2018. A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology* 31(1), 20 (2018), 1–23. https://doi.org/10.5334/irsp.181

[87] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems - RecSys 11*. 157–164.

[88] Silvia Quarteroni and Suresh Manandhar. 2008. Designing an interactive open-domain question answering system. *Natural Language Engineering* 15, 1 (2008), 73–95. https://doi.org/10.1017/S1351324908004919

[89] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, and Francesca Toni. 2020. Argumentation as a Framework for Interactive Explanations for Recommendations. In *Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning*. 805–815.

[90] Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 583–593.

[91] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and Metrics for Cold-Start Recommendations. In *Proceedings of SIGIR 2002*. 253–260.

[92] Wolfgang Schnotz. 2014. Integrated Model of Text and Picture Comprehension. In *The Cambridge Handbook of Multimedia Learning (2nd ed.).* 72–103.

[93] Kacper Sokol and Peter Flach. 2020. LIMEtree: Interactively Customisable Explanations Based on Local Surrogate Multi-output Regression Trees. (2020).

[94] Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. 34, 2 (2020), 235–250.

[95] Ji Hee Song and George M. Zinkhan. 2008. Determinants of Perceived Web Site Interactivity. *Journal of Marketing* 72, 2 (2008), 99–113.

[96] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook.* Springer US, Boston, MA, 353–382.

[97] Stephen E. Toulmin. 1958. The Uses of Argument. (1958).

[98] Suzan Verberne, Maarten van der Heijden, Max Hinne, Maya Sappelli, Saskia Koldijk, Eduard Hoenkamp, and Wessel Kraaij. 2013. Reliability and validity of query intent assessments: Reliability and Validity of Query Intent Assessments. *Journal of the American Society for Information Science and Technology* 64, 11 (2013), 2224–2237.

[99] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent User Interfaces.* ACM, 47–56.

[100] Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *15th International Conference on Intelligent Text Processing and Computational Linguistics.* 115–127.

[101] Douglas Walton. 2000. The Place of Dialogue Theory in Logic, Computer Science and Communication Studies. 123 (2000), 327–346.

[102] Douglas Walton. 2004. A new dialectical theory of explanation. 7, 1 (2004), 71–89.

[103] Douglas Walton. 2011. A dialogue system specification for explanation. 182, 3 (2011), 349–374.

[104] Douglas Walton and Erik C. W. Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning.* State University of New York Press, New York.

[105] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 18.* 165–174.

[106] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. 62, 6 (2019), 70–79.

[107] Yao Wu and Martin Ester. 2015. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Eighth ACM International Conference on Web Search and Data Mining.* ACM, 153–162.

[108] Bo Xiao and Izak Benbasat. 2007. ECommerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly* 31, 1 (2007), 137–209.

[109] Markus Zanker and Martin Schoberegger. 2014. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems.* 33–36.

[110] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 177–186. https://doi.org/10.1145/3269206.3271776

[111] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval.* 83–92.

A CONVEX-DS COLLECTION AND ANNOTATION ARTIFACTS

A.1 WoOz pre-study

We used elaborated the following scheme (Fig. 15), as the guideline for the wizard (Section 6.1), aiming to portray a structured conversation similar across participants.

A.2 Corpus collection: Annotation guidelines

Figures 16,17 and 18 depict the guidelines presented to the annotators for the process of annotating the corpus described in Section 6.2.

B USER EVALUATION QUESTIONNAIRES

Tables 7 and 8 show the questionnaire items used in the user study reported in section 7.

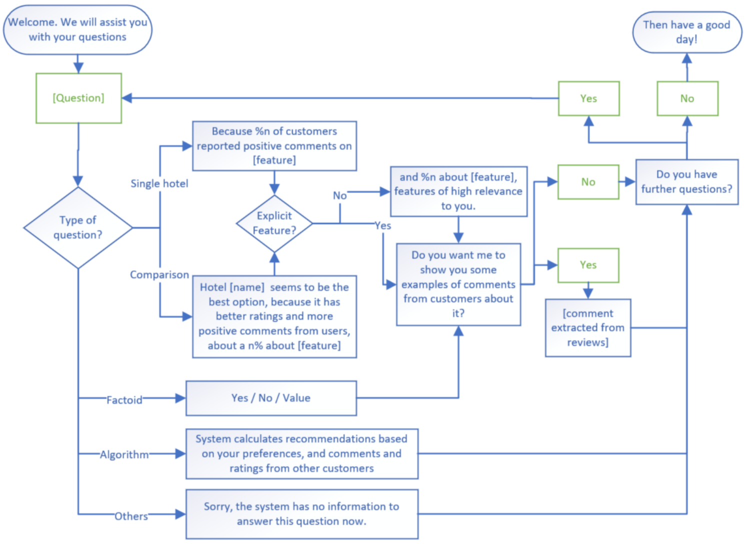
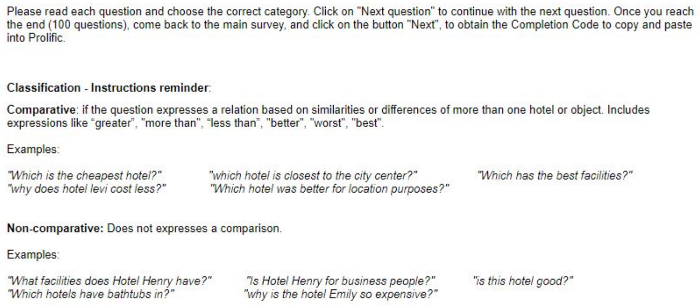

Welcome. We will assist you with your questions
Then have a good day!
[Question]
Yes
No
Because %n of customers reported positive comments on [feature]
and %n about [feature], features of high relevance to you.
No
Do you have further questions?
Single hotel
Type of question?
Explicit Feature?
No
Yes
Do you want me to show you some examples of comments from customers about it?
Yes
Comparison
Hotel [name] seems to be the best option, because it has better ratings and more positive comments from users, about a n% about [feature]
[comment extracted from reviews]
Factoid
Yes / No / Value
Algorithm
System calculates recommendations based on your preferences, and comments and ratings from other customers
Others
Sorry, the system has no information to answer this question now.

Fig. 15. Wizard guideline for conversational explanations used in WoOz experiment (Section 6.1), as reported in [43]. Blue boxes represent utterances by the system, green boxes the utterances by users.



Please read each question and choose the correct category. Click on "Next question" to continue with the next question. Once you reach the end (100 questions), come back to the main survey, and click on the button "Next", to obtain the Completion Code to copy and paste into Prolific.

Classification - Instructions reminder:

Comparative: if the question expresses a relation based on similarities or differences of more than one hotel or object. Includes expressions like "greater", "more than", "less than", "better", "worst", "best".

Examples:

"Which is the cheapest hotel?"
"why does hotel levi cost less?"

"which hotel is closest to the city center?"
"Which hotel was better for location purposes?"

"Which has the best facilities?"

Non-comparative: Does not expresses a comparison.

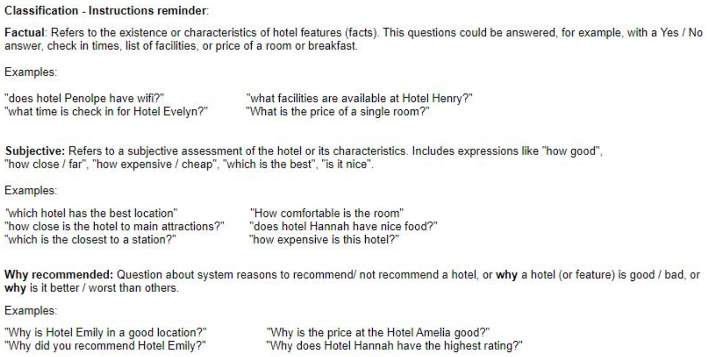
Examples:

"What facilities does Hotel Henry have?"
"Which hotels have bathtubs in?"

"Is Hotel Henry for business people?"
"why is the hotel Emily so expensive?"

"is this hotel good?"

Fig. 16. Annotation guideline, dimension comparison.



Classification - Instructions reminder:

Factual: Refers to the existence or characteristics of hotel features (facts). This questions could be answered, for example, with a Yes / No answer, check in times, list of facilities, or price of a room or breakfast.

Examples:

"does hotel Penolpe have wifi?"
"what time is check in for Hotel Evelyn?"
"what facilities are available at Hotel Henry?"
"What is the price of a single room?"

Subjective: Refers to a subjective assessment of the hotel or its characteristics. Includes expressions like "how good", "how close / far", "how expensive / cheap", "which is the best", "is it nice".

Examples:

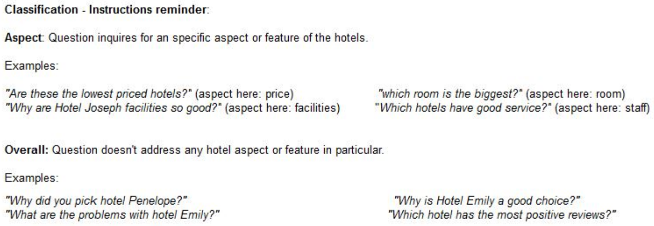
"which hotel has the best location"
"how close is the hotel to main attractions?"
"which is the closest to a station?"
"How comfortable is the room"
"does hotel Hannah have nice food?"
"how expensive is this hotel?"

Why recommended: Question about system reasons to recommend/ not recommend a hotel, or **why** a hotel (or feature) is good / bad, or **why** is it better / worst than others.

Examples:

"Why is Hotel Emily in a good location?"
"Why did you recommend Hotel Emily?"
"Why is the price at the Hotel Amelia good?"
"Why does Hotel Hannah have the highest rating?"

Fig. 17. Annotation guideline, dimension assessment.



Classification - Instructions reminder:

Aspect: Question inquires for an specific aspect or feature of the hotels.

Examples:

"Are these the lowest priced hotels?" (aspect here: price)
"Why are Hotel Joseph facilities so good?" (aspect here: facilities)
"which room is the biggest?" (aspect here: room)
"Which hotels have good service?" (aspect here: staff)

Overall: Question doesn't address any hotel aspect or feature in particular.

Examples:

"Why did you pick hotel Penelope?"
"What are the problems with hotel Emily?"
"Why is Hotel Emily a good choice?"
"Which hotel has the most positive reviews?"

Fig. 18. Annotation guideline, dimension detail.

Table 7. Item questions used in user study reported in Sec. 7

Code	**Question**	**Analisys Variable**	**Reference**
AI03	What are the 5 most important aspects for you when looking for a hotel?	sort_aspects	
UT02_01	I prefer to gather all the necessary information before committing to a decision.	UT_rational	[37]
UT02_02	I thoroughly evaluate decision alternatives before making a final choice.	UT_rational	[37]
UT02_03	In decision making, I take time to contemplate the pros/cons or risks/benefits of a situation.	UT_rational	[37]
UT02_04	Investigating the facts is an important part of my decision-making process.	UT_rational	[37]
UT02_05	I weigh a number of different factors when making decisions.	UT_rational	[37]
UT02_06	When making decisions, I rely mainly on my gut feelings.	UT_intuitive	[37]
UT02_07	My initial hunch about decisions is generally what I follow.	UT_intuitive	[37]
UT02_08	I make decisions based on intuition.	UT_intuitive	[37]
UT02_09	I rely on my first impressions when making decisions.	UT_intuitive	[37]
UT02_10	I weigh feelings more than analysis in making decisions.	UT_intuitive	[37]
UT02_18	I am competent when it comes to graphing and tabulating data.	UT_visual_fam	[56]
UT02_19	I frequently tabulate data with computer software.	UT_visual_fam	[56]
UT02_20	I have graphed a lot of data in the past.	UT_visual_fam	[56]
UT02_21	I frequently analyze data visualizations.	UT_visual_fam	[56]
EE02_06	I would recommend the system to others.	sys_effectiveness	[55]
EE02_07	I could make better decisions with the help of the system.	sys_effectiveness	[55]
EE02_08	The recommender is useful.	sys_effectiveness	[55]
EE02_10	I believe that the recommender would act in my best interest.	sys_trust	[74]
EE02_11	I would characterize recommender as honest.	sys_trust	[74]
EE02_12	Recommender is competent and effective in providing hotel recommendations.	sys_trust	[74]
EE02_13	I would feel comfortable depending on the information provided by recommender.	sys_trust	[74]
EE02_14	I would want to use recommender again.	sys_trust	[74]
EE02_15	I would confidently book a hotel based on recommendation I was given by recommender.	sys_trust	[74]
EE02_16	I would be willing to share the specifics of my preferences when looking for a hotel to the recommender.	sys_trust	[74]
EE05_04	The explanations make me confident that I would like the hotel I chose.	expl_quality (expl_confidence)	[56]
EE05_05	The explanations make the recommendation process clear to me.	expl_quality (expl_transparency)	[56]
EE05_06	I would enjoy using a recommender system if it presented explanations in this way.	expl_quality (expl_satisfaction)	[56]
EE05_08	Explanations were convincing.	expl_quality (expl_persuasiveness)	[56]
EE05_20	The explanations provided sufficient information to make my decision.	expl_quality (expl_sufficiency)	[28]
EE03_01	Please let us know about your overall opinion about the explanations provided or how they could be improved:	comments	
HC03_01	Please provide the reasons why you chose the hotel you did.	reasons	
EE07_01	How would you explain to a friend, in your own words, how the system generates recommendations?.	explain_to_someone	

Table 8. Factor loadings of questionnaire items to evaluate perception of RS transparency, used in user study reported in Sec. 7

Constructs	Code	Items	Estimate	Stand. Error	Z	p	Stand. Estimate
Input	EE02_22	It was clear to me what kind of data the system uses to generate recommendations.	0.589	0.069	8.5	< .001	0.653
	EE02_23	I understood what data was used by the system to infer my preferences.	0.641	0.065	9.93	< .001	0.746
	EE02_24	I understood which item characteristics were considered to generate recommendations.	0.425	0.059	7.22	< .001	0.567
Output	EE02_25	I understood why the items were recommended to me.	0.497	0.070	7.11	< .001	0.612
	EE02_26	I understood why the system determined that the recommended items would suit me.	0.466	0.067	6.93	< .001	0.593
	EE02_27	I can tell how well the recommendations match my preferences.	0.504	0.069	7.29	< .001	0.573
Functionality	EE02_28	The system provided information to understand why the items were recommended.	0.459	0.075	6.12	< .001	0.502
	EE02_29	The system provided information about how the quality of the items was determined.	0.506	0.073	6.92	< .001	0.549
	EE02_30	The system provided information about how my preferences were inferred.	0.475	0.068	6.98	< .001	0.550
	EE02_31	The system provided information about how well the recommendations match my preferences.	0.519	0.065	8.01	< .001	0.616
	EE02_32	I understood how the quality of the items was determined by the system.	0.542	0.057	9.56	< .001	0.709
Interaction	EE02_33	I know what actions to perform in the system so that it generates better recommendations	0.918	0.130	7.08	< .001	0.935
	EE02_34	I know what needs to be changed in order to get better recommendations	0.513	0.094	5.44	< .001	0.543

DOI: 10.17185/duepublico/75833
URN: urn:nbn:de:hbz:465-20220428-111818-5