Assessing the Helpfulness of Review Content for Explaining Recommendations

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

Despite the maturity already achieved by recommender systems algorithms, little is known about how to obtain and provide users with a proper rationale for a recommendation. Transparency and effectiveness of recommender systems may be increased when explanations are provided. In particular, identifying of helpful argumentative content from reviews can be leveraged to generate textual explanations. In this paper, we investigate the reasons why a review might be considered helpful, and show that the perception of credibility and convincingness mediates the relationship between helpfulness and the perception of objectivity and relevant aspects addressed. Our findings led us to suggest an argument-based approach to automatically extracting helpful content from hotel reviews, a domain that differs from those that best fit classical argumentation theories.

CCS CONCEPTS

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

KEYWORDS

Recommender systems, user study, explanations, argument mining

ACM Reference Format:
Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2019. Assessing the Helpfulness of Review Content for Explaining Recommendations. In *Proceedings of SIGIR 2019 Workshop on ExplainAble Recommendation and Search (EARS'19).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND RELATED WORK

Recommender Systems (RS) have become widespread tools that aim at facilitating users' search and decision making in a large variety of domains such as e-commerce or media streaming. . While the algorithmic accuracy of RS has been improved considerably over the years, from a user perspective, most systems act as black boxes that provide no rationale for their decisions and that do not allow users to control or question the recommendations given [9]. Research in RS [19] has shown that providing explanations can be an effective means for increasing transparency, facilitating users' decision-making and increasing user satisfaction as well as the trustworthiness of the system . In particular, textual explanations can provide valuable information that helps users to make better decisions [6]. In this regard, the interest in exploiting online reviews in RS has increased in recent years, given the richness of its content on aspects of interest to the user, not only to improve the accuracy of RS predictions [3, 11, 14, 23], but also to provide textual explanations of the recommended items [6, 22]. For example, [22] proposed a matrix factorization based model to leverage sentiment analysis on aspects addressed in reviews, to provide both recommendations and textual explanations based on templates. [23] proposed a deep learning approach, using two parallel neural networks to simultaneously model items and users from reviews, in order to predict ratings. This approach was extended by [4], who stressed that not all reviews should contribute in the same way to item modeling, and that predictions could be improved by using reviews that were useful to the user. For this purpose, an attention mechanism was used, demonstrating that this could not only improve the accuracy of predictions compared to approaches that consider all reviews equally, but also select relevant information that could be used to provide explanations. Despite the convenience of such a concept, [4] determines usefulness by only addressing user's aspects of interest. In contrast, there are many additional features that can also contribute to usefulness.

Extensive research on review helpfulness prediction has been highly focused on content features (eg. review length, readability or word-based features) [10, 13, 21]. However, helpfulness could also be predicted by using features that convey convincing content and provide hints to credibility. In this sense, and given its persuasive nature, an argument-based approach could help to identify this type of content, through the understanding of underlying patterns and the argumentative-type features contained in reviews, bearing in mind the impact that appropriate argumentation has on the beliefs and behaviors of consumers towards an item. Therefore, we believe that the exploiting of arguments given by users in reviews can contribute to helpfulness prediction. Nevertheless, little is known about whether features related to arguments can work as a good predictors of the helpfulness. In this regard, [12] coined the term "argument-based features" (e.g ratio between claims and premises or position of the argument components), examined and annotated argument components in reviews of hotels domain, and found that when argument-based features are used, helpfulness prediction performance is increased over the use of traditional content features. Despite these findings, it is important to note that there is still a lack of consensus on whether the reviews reflect argumentative structures, not to mention the lack of reliable annotated corpus that would facilitate the automatic extraction of such a content [8], not

EARS'19, July 25, 2019, Paris, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

being the case for domains such as legal documents and debates, in which computational argument mining has made significant progress in recent years.

Overall, we describe in this paper our approaches to explainable recommendations, focusing on the extraction of arguments from reviews. We hypothesise that the use of helpful reviews with convincing and credible arguments that refer to aspects of interest to the user can improve not only rating predictions, but also be the base to generate useful textual explanations. As a first step, we seek to assess in this paper what makes a review be perceived as helpful. Accordingly, we present a user study in which users were asked to rate the helpfulness of a series of reviews and to report their reasons for voting. Here, our specific hypothesis is that features related to arguments (i.e. balance between pros and cons, opinions supported by facts, and a stringent flow of arguments) can work as good predictors of helpfulness in the hotel reviews domain, as well as coverage of aspects relevant to users, and even better than other traditional content features, i.e. review length and the level of detail. Lastly, we propose an argument-based approach to extracting helpful content from reviews, based on our findings.

2 HELPFULNESS STUDY

We hypothesized that features related to arguments can work as good predictors of helpfulness in the hotel reviews domain, in addition to the addressing of aspects relevant to users, and better than other traditional content features, i.e. review length and level of detail.

2.1 Method

We conducted a study survey with 108 participants (49 female, mean age 34.38 and range between 19 and 68) recruited through Amazon Mechanical Turk. All participants were requested to read and rate helpfulness for the same set of 18 hotel reviews.

Procedure: We randomly selected 18 reviews for three hotels in the San Francisco area. For this, we used the ArguAna dataset for argumentation analysis in hotels domain [20], which includes annotations of positive and negative opinions, neutral statements, and aspects addressed. We asked the participants to read 6 reviews for each of the 3 selected hotels, presented at random, and to rate them according to how helpful they considered the review was, and the reasons that contributed to that perception. We asked validation questions at the end of each set of hotel reviews, in order to ensure that participants were actively reading the content. No information of the overall sentiment rate given by author or further information about her was given, seeking for participants to only evaluate the text contained in the review.

Questionnaire: The review text was displayed, followed by the question "How helpful was this review?" and a 1–5 Likert-scale for the response (1: Not helpful at all, 5: Very helpful). Next, participants were requested to inform their opinion about 10 different reasons for their helpfulness reply, in relation to: 1) argument-based features, which are the focus of this study: *objectivity* ("the review includes an adequate amount of objective statements based on facts"), *pros and cons* ("the review provided a balanced view of pros and cons") and *flow of arguments* ("the review has a stringent flow of arguments"); 2) content-based features: *length* ("the review was too short or too long") and *level of detail* ("the level of detail provided was too little / too much"); 3) *aspects* ("the review addresses the aspects that are relevant for my purposes"); 4) perception of *credibility* ("the review seems credible") and *convincingness* ("the review provided convincing reasons"); 5) content that is *emotional* ("the review contains emotional content") or *episodic* ("the review contains information that might be only episodic"). Opinions were rated on a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

2.2 Results and Discussion

Multiple logistic regression analysis was used to test the reasons that significantly predict participants' ratings of helpfulness. Here, positive coefficients would indicate that the reason contributes to an increase of helpfulness perception, whereas negative coefficients would indicate the opposite effect. The results of the regression indicated the 10 predictors explained 25.4% of the variance and the model predicts 74.5% of the responses correctly. The Pseudo-R^2 is 0.23. As expected, credibility significantly predicted helpfulness (β= .29, p<.001), as did aspects. Results also revealed that an increase in the variable length leads to a decrease in the perception of helpfulness (β=-.20, p<.01). In contrast to [21], this finding confirms our expectation that long reviews are not necessarily considered helpful, and that other reasons related to argumentative content can have a positive impact on the perception of helpfulness, at least in hotels domain. Such is the case with the variables objectivity (β=.23, p<.01) and convincingness (β=.23, p<.01). In the other hand, contrary to our expectations and previous findings [5], pros and cons is not strongly correlated to helpfulness , nor is flow of arguments. The latter might suggest that even if helpful reviews involve some sort of argumentative content, they do not necessarily follow a strict flow of arguments; in other words, it could be inferred that the presentation of a list of arguments is sufficient, and that a stringent flow of arguments is not really relevant to the perception of helpfulness in this domain. On the other hand, the level of detail not only is not a significant predictor of helpfulness, but we found that is highly correlated to length (r_s=.69, p<0.001), reason why we decided to exclude it from further analyses. Additionally, we found that emotional is also not a significant predictor of helpfulness; even when some of the reviews than can be read online appeal to emotions in order to persuade, this may be a dimension that does not apply transversely to most types of users, as if it could be the case of objectivity. Lastly, we consider that, even when episodic is not a significant predictor of helpfulness, is an interesting aspect that should be addressed further. We consider that content related to an incident that could be perceived as irrelevant in a single review -but appears repeatedly in subsequent reviews- could change the perception of helpfulness, once several reviews are read. Table 1 shows the coefficients obtained for all the model variables.

In addition, we wanted to understand in more detail the mechanisms underlying the relationship between helpfulness and observed variables that can be considered 'less subjective' (in contrast to variables like credibility and convincingness), since they can be quantified through the use of complementary information. (e.g. participants were asked to inform their aspects of interest, and reviews are already annotated with aspects, facts and opinions).

Table 1: Logistic regression of reasons for voting on helpfulness

Reason	Coefficient[a]	p value	Odds ratio
Length	-0.200**	0.00	0.82
Level of detail	-0.108	0.11	0.90
Objectivity	0.232**	0.00	1.26
Pros and cons	0.007	0.91	1.01
Convincing	0.225**	0.00	1.25
Aspects	0.228**	0.00	1.26
Flow of Arguments	0.024	0.71	1.02
Credibility	0.293***	0.00	1.34
Emotional	0.048	0.42	1.05
Episodic	-0.074	0.22	0.93
Accuracy:	74.5%		
Pseudo-R^2:	0.23		

[a](*) p<0.05, (**) p<0.01, (***) p<0.001

Credibility and convincingness could, on the other hand, be evaluated as mediators of the relationship between helpfulness and some of the rest of observed variables. To this respect, two mediation models were tested in three steps according to standard procedure [2]. First, we regressed helpfulness on all independent variables but credibility and convincingness, and we confirmed that the length (β=-.18, p<0.01), objectivity (β=.28, p<0.001) and aspects (β=.31, p<0.001) are significant predictors of helpfulness. Second, we regressed the mediator credibility on the independent variables, and confirmed that objectivity (β=.14, p<0.05) and aspects (β=.18, p< 0.01) are significant predictors of the credibility. The same was the case with convincing as mediator, and confirmed that details (β=-.19, p<0.01), as well as objectivity (β=.17, p<0.05) and aspects (β=.39, p<0.001) are significant predictors of convincingness. In the third step, the association of aspects and objective and helpfulness reduced significantly when credibility and convincingness were added to the model. Therefore, credibility and convincingness act as mediators of the relationship between helpfulness and objectivity and relevant aspects addressed (Figure 1). The foregoing means that credibility and convincingness serve to clarify the nature of the relationship between helpfulness and aspects and objectivity. Thus, addressing relevant aspects and objectivity influences credibility and convincing, which in turn influence helpfulness.

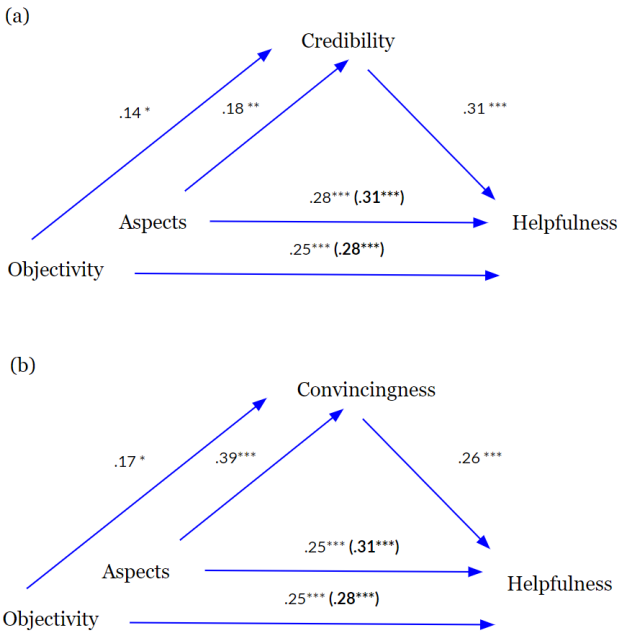

Figure 1: Credibility (a) and convincingness (b) as mediators between helpfulness and aspects and objectivity. Coefficients in bold represent the strength of linear relation between helpfulness and aspects and objectivity, when credibility and convincing are not statistically controlled by including them as predictors of helpfulness.

3 AN ARGUMENT-BASED APPROACH TO EXTRACTING HELPFUL CONTENT FROM REVIEWS

Even when classical argumentative structures are not representative of the content normally found in online reviews, the authors offer, in many cases, valuable information of persuasive purpose. In this sense, and based on the findings of our study, our proposed approach is based on the idea that objective statements supported by facts provide very helpful information to users. In consequence, helpful content could be extracted by means of a shallow argumentative structure, that represents objective reasons or evidences that support an opinion on a certain aspect. A similar approach is used by [17], where an argument is represented by 'reasons supporting a conclusion', although no distinction is made between objective and subjective reasons. On the other hand, we define objective reasons as statements that include facts, and subjective opinions as statements with polarity in respect to an aspect, e.g. "Don't count on the breakfast that is included [opinion, polarity: negative]; instant coffee and packaged muffins, etc [fact]" For the automatic extraction of these components, we plan to use the ArguAna corpus [20], which includes a statement-level annotation of facts, positive and negative opinions, and aspects for hotel reviews extracted from TripAdvisor (inter-annotator agreement Fleiss' κ=0.67). In spite of the convenience of this corpus, the annotation scheme needs to be complemented in order to useful for our purposes. In particular, the annotation of facts that relate to properties and services provided by the hotel is still needed, since neutral statements from ArguAna include not only facts, but all the statements without polarity, like contextual information (e.g. "We stayed here New Year's Eve"), or suggestions and tips that are not directly related to the evaluated hotel (e.g. "We found a great Japanese restaurant called Wasabi and Ginger on Van Ness Avenue just 3 mins walk away - highly recommended"). In addition, the support relation between opinions and facts should also be annotated, as a basis for the automatic relation detection task.

Table 2: Examples of additional argument figures found in reviews perceived as helpful

Figure	Aim	Example
Match reality vs. expectation	Establish a contrast or a match between initial expectation and the real perception of author.	"you get what you pay for."
Moderately negative/ positive statements	Establish nuances of polarity, which may reflect a honest author's opinion, and consequently increase the perception of credibility.	"A low-cost place to sleep with no frills."
Episodic by example	Enumerate examples to support an opinion. In this case, unlike facts like breakfast is included, we refer to events that could be episodic.	"The staff was very friendly and helpful. For example, they set up transportation for me on multiple days without a hitch."
Preferential	Ilustrate opinions that reflect author's personal preferences, that involve a comparison with similar items.	"Room was clean and the place was nice. But I enjoyed some of the other Joie De VIvre hotels in the same two block area more."

In addition, we examined the content of reviews with the best helpfulness scores in our user study (section 2), and found a number of argumentative figures that represent potentially convincing content (Table 2). In the future, a detailed analysis is required to establish their impact on helpfulness, the feasibility of its annotation and automatic extraction, and to what extent its use would improve the quality of the extracted arguments.

We present above our approach to extracting helpful arguments from reviews. However, our aim is, in a broader sense, to use such content not only to improve rating predictions, but also to generate helpful textual explanations that reflect system transparency and facilitate users decision making. Therefore, in future work, we plan to develop and integrate methods that:

- Facilitate the automatic extraction of helpful arguments from reviews. We plan to base our work on developments from argument mining [7, 8, 18] (for the supervised detection of units and its relationships on a multi-sentence level), sentiment and subjectivity analysis (for detection of objective/subjective units on a sentence level [16]).
- Establish the relevance of the extracted arguments to generate proper explanations and improve rating prediction, by leveraging attention mechanisms [1] to detect aspects of interest to user.
- Allow to generate natural language explanations of recommended items, using the relevant arguments extracted. Here, we plan to base our work on abstractive summarization techniques as proposed by [15].

According to our initial approach, all reviews containing arguments (subjective opinions supported by objective facts) that are relevant to users' interests would be considered helpful and used as the basis for RS predictions and explanations. This binary approach (review is helpful or not) could be extended to determine different degrees of helpfulness, so that higher quality arguments have a greater weight in the rating prediction. In this respect, the definition of additional criteria to improve the evaluation of the relative quality of arguments is still needed; however, the use of features such as the ratio between facts and opinions, or the length of sentences that address aspects relevant to users could be a starting point for this purpose.

Furthermore, we also plan to address in the future the applicability of the proposed approach to domains other than hotel reviews, e.g. restaurant or product reviews.

4 CONCLUSIONS

In this paper we have discussed the use of online reviews, in particular to generate explanations that can serve the objectives of transparency and effectiveness of RS. A novel way of extracting helpful content from reviews was proposed, based on a arguments-based approach. To support our proposal, we presented a user study, which sought to establish the reasons why a user might find a review helpful. Our findings lead us to suggest that arguments in the form of opinions supported by objective statements provide very helpful information to users. As future work, we aim to implement this concept and use it, not only to improve RS predictions, but also to generate textual explanations of the recommended items.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, , and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *arXiv preprint arXiv:1409.0473*.

[2] Reuben M. Baron and David A. Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51 (1986), 1173–1182.

[3] Rose Catherine and William Cohen. 2017. TransNets: Learning to Transform for Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 288–296.

[4] 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.

[5] Laura Connors, Susan M. Mudambi, and David Schuff. 2011. Is It the Review or the Reviewer? A Multi-Method Approach to Determine the Antecedents of Online Review Helpfulness. In *Proceedings of the 44th Hawaii International Conference on System Sciences*. 1530–1605.

[6] 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.

[7] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 11– 22.

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

[9] 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.

[10] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. 423–430.

[11] Guang Ling, Michael R. Lyu, and Irwin King. 2014. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender systems*. 105–112.

[12] Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using Argument-based Features to Predict and Analyse Review Helpfulness. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1358–1363.

[13] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 334–342.

[14] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.

[15] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 280–290.

[16] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. In *arXiv preprintarXiv:1704.01444*.

[17] Patrick Saint-Dizier. 2012. Processing natural language arguments with the TextCoop platform. *Argument and Computation* 3 (2012), 49–82.

[18] Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 35–41.

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

[20] 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.

[21] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 38–44.

[22] 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.

[23] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 425–434.