User and Item-aware Estimation of Review Helpfulness

Noemi Mauro[a], Liliana Ardissono[a,*], Giovanna Petrone[a]

[a]*Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, I-10149 Torino, Italy*

Abstract

In online review sites, the analysis of user feedback for assessing its helpfulness for decision-making is usually carried out by locally studying the properties of individual reviews. However, global properties should be considered as well to precisely evaluate the quality of user feedback.

In this paper we investigate the role of *deviations* in the properties of reviews as helpfulness determinants with the intuition that "out of the core" feedback helps item evaluation. We propose a novel helpfulness estimation model that extends previous ones with the analysis of *deviations in rating, length and polarity* with respect to the reviews written by the same person, or concerning the same item. A regression analysis carried out on two large datasets of reviews extracted from Yelp social network shows that user-based deviations in review length and rating clearly influence perceived helpfulness. Moreover, an experiment on the same datasets shows that the integration of our helpfulness estimation model improves the performance of a collaborative recommender system by enhancing the selection of high-quality data for rating estimation. Our model is thus an effective tool to select relevant user feedback for decision-making.

Keywords: Review helpfulness, helpfulness determinants, regression analysis, helpfulness-aware personalized item recommendation

Declarations of interest: none.

*This is to indicate the corresponding author.

Email addresses: noemi.mauro@unito.it (Noemi Mauro), liliana.ardissono@unito.it (Liliana Ardissono), giovanna.petrone@unito.it (Giovanna Petrone)

Published in Information Processing & Management, Elsevier.
DOI: https://doi.org/10.1016/j.ipm.2020.102434.
Link to the page of the paper on Elsevier web site:
https://www.sciencedirect.com/science/article/pii/S0306457320309274

1. Introduction

Reviews of items posted in e-commerce sites and social media are a precious source of information about consumers' experience with products but their abundance challenges their effective fruition. Marketers thus attempt to evaluate the helpfulness of reviews in order to promote those which best support purchasing decisions. However, depending on factors such as age (recentness) and level of visibility in the web sites, good reviews might fail to get feedback from readers (Hu & Chen, 2016; Hu et al., 2017). Therefore, helpfulness must be estimated *a priori* to sort comments in an informed way as soon as they are posted (Krestel & Dokoohaki, 2015).

Several researchers assume that helpfulness is an internal property of reviews. For instance, see Mudambi & Schuff (2010), Yang et al. (2015), Hong et al. (2017) and Siering et al. (2018). In these works, each comment is analyzed independently of the others. However, linguistic style is personal (Li et al., 2019) and perceived helpfulness also depends on the variability of ratings provided by reviewers (Gao et al., 2017). Moreover, Raghavan et al. (2012) and Fang et al. (2016) observed that the deviation with respect to the mean rating of a product supports helpfulness estimation. Starting from these findings, we

are interested in understanding whether a contextual analysis of reviews written by the same person, or concerning the same item, contributes to enhance helpfulness assessment. Specifically, we investigate the role of deviations in the content properties of reviews as helpfulness determinants. We pose the following research questions:

RQ1: *Given a review r, does a deviation from the mean length, polarity and rating of the other reviews written by the same person provide useful information to assess the perceived helpfulness of r?*

RQ2: *Given a review r, does a deviation from the mean length, polarity and rating of the other reviews on the same item provide useful information to assess the perceived helpfulness of r?*

In order to answer these questions we analyzed two datasets of reviews from Yelp (2019a), one about accommodation services (including about 10000 comments) and the other about food services (including about 65000 comments). We learned through regression a helpfulness estimation model that combines (i) largely used determinants such as review length, TF/IDF statistics and ratings with (ii) review polarity, that is the sentiment emerging from its text, and (iii) the deviations of these features among the comments provided by the same user, or concerning a single item. Then, we compared the helpfulness estimation capability of our model with that of baseline models that only use factors of type (i) and (ii). We carried out the evaluation as follows:

1. First, we checked whether our model estimates helpfulness more accurately than the baselines by correlating the predicted values with the feedback about reviews observed in the datasets (ground-truth helpfulness) by means of Pearson and Spearman analyses.
2. Then, we evaluated whether our model supports the identification of high-quality information for decision-making. We did this by extending a collaborative recommender system (Koren & Bell, 2011) to weight the impact of observed rating data on recommendation, and by comparing suggestion performance with that of standard Collaborative Filtering.

On both datasets the experimental results show that our model better adheres to ground-truth helpfulness than the baselines. Moreover, our model enhances recommendation performance in terms of accuracy, error minimization and ranking of items. In summary, we provide the following contributions:

- An advancement of the state of the art in review helpfulness estimation based on the idea that "out of the core" reviews can be relevant information sources for decision-making.
- A novel helpfulness prediction model that extends previous ones with the identification of user-based and item-based helpfulness determinants.
- An experimental validation which shows that our model outperforms the selected baselines and improves collaborative item recommendation.

Our work paves the way toward the development of human-centered algorithms by enhancing performance and transparency of recommender systems. Specifically, review helpfulness prediction can be used to select high-quality ratings for recommendation. Moreover, it can be employed to explain the suggestions generated by the system using the textual feedback provided by previous consumers to describe their experience with items (Ghose & Ipeirotis, 2011).

The remainder of this article is organized as follows. Section 2 provides a literature review. Section 3 presents our research methodology and Section 4 describes the experimental results. Section 5 shows the benefits of our model to personalized item recommendation. Section 6 summarizes the findings of our study and their implications. Section 7 describes limitations and suggestions for future work and Section 8 concludes the paper.

2. Background and Related Work

This section positions our work with respect to the research about helpfulness determinants concerning review content. See Ocampo Diaz & Ng (2018) for a survey. Table 1 summarizes these factors by reporting, for each one, the impact on perceived helpfulness and the works supporting the finding. Most of the cited

Table 1: Determinants of perceived review helpfulness concerning review content.

Determinant	Definition	Prior finding	Represented study
length (depth)	total number of words in a review	positive, negative, inverted-U-shaped	Kim et al. (2006), Wu (2017), Mudambi & Schuff (2010), Fink et al. (2018), Eslami et al. (2018) Hong et al. (2017) Raghavan et al. (2012)
Unigram	TF/IDF value of the words included in the review	not specified	Kim et al. (2006), O'Mahony & Smyth (2018)
entropy	entropy of the words included in the review	negative	Fresneda & Gefen (2019)
rating	star rating in [1, 5]	positive, inverted-U-shaped	Kim et al. (2006), Eslami et al. (2018), Mudambi & Schuff (2010), O'Mahony & Smyth (2010, 2018)
writing style	readability, linguistic correctness	positive, domain dependent, insignificant	Ghose & Ipeirotis (2011), Liu et al. (2019) Hong et al. (2017), Krishnamoorthy (2015)
subjectivity	subjective statements in review	mix is negative	Ghose & Ipeirotis (2011), Krishnamoorthy (2015)
linguistic features	adjectives; state and action verbs; ...	positive	Krishnamoorthy (2015)
semantic features	total number of concepts in the review; average number of concepts per sentence	positive	Qazi et al. (2016), Cao et al. (2011) Sun et al. (2019)
polarity (valence, sentiment)	positive/negative sentiment of review	negative	Eslami et al. (2018), Dong et al. (2013), Siering et al. (2018), Salehan & Kim (2016)
aspects	semantic features occurring in reviews	depends on polarity	Paul et al. (2017), Xiong & Litman (2014), Yang et al. (2015, 2016)
coherence (consistency)	consistency between review polarity and rating, similarity between review title and content	–, negative	Dong et al. (2013), Zhou et al. (2020), Shen et al. (2019)
rating deviation	deviation of rating from mean product rating	positive	Raghavan et al. (2012)
domain-specific content	domain-specific item properties	moderated by product	Ahmad & Laroche (2017)
age	number of days since publication	positive	Hong et al. (2017), (Hu & Chen, 2016)

works study a larger set of determinants but Table 1 only shows those identified as influential by the authors of the cited works.

2.1. Review helpfulness for item recommendation

Some researchers use reviews to calculate the associated ratings by means of a content analysis. For instance, see Margaris et al. (2020). Our work is different because we measure review helpfulness to *select* high-quality ratings for recommendation. This selection is a pre-requisite for the generation of relevant suggestions because algorithmic accuracy is not sufficient when working with poor data. Moreover, this selection supports recommender systems transparency by identifying appropriate feedback about items that can be used to explain the generated results. This improves traceability and trust which, as discussed by Shin (2020a), enhance user acceptance of services. Specifically, Shin et al. (2020) have discovered that algorithmic experience "is inherently related to human understanding of fairness, transparency, and other conventional components of user-experience" which, in turn, are tightly connected to explainability.

Recommender systems (Ricci et al., 2011) address transparency (Tintarev & Masthoff, 2015) and trust (Berkovsky et al., 2017, 2018) by enriching the suggestions they generate with a description of the degree, or of the type of matching between users and items. For instance, see Herlocker et al. (2000), Kouki et al. (2019) and Pu & Chen (2007). We claim that item reviews perceived as helpful by their readers are an important asset to be used for this purpose because they make it possible to describe item properties by exploiting previous consumer experience (Mauro et al., 2020a; Ghose & Ipeirotis, 2011). This is in line with findings related to the news recommendation domain, in which Shin (2020b) has pursued interactivity and presented "a news recommendation experience model incorporating algorithm quality (transparency and accuracy) and perceived value (utility and convenience) as antecedent factors of confirmation and satisfaction."

2.2. Review-related helpfulness determinants

2.2.1. Structural features

Review length, rating (number of stars) and Unigram (TF/IDF statistics of the words appearing in the review (Robertson, 2004)) are recognized as important helpfulness determinants as found by Kim et al. (2006). Length is taken as a proxy of informativeness and can be associated to user involvement in writing the comment (Pan & Zhang, 2011). Fink et al. (2018) observed that, when content is created under low-constraint settings, such as mobile interaction, length has an inverted-U-shaped influence on perceived helpfulness, with medium length comments being more effective than very short and very long ones. Rating is taken as a proxy of review valence representing positive/negative opinion. Unigram assesses the relevance of review words when compared to the other comments about the same product. It can be noticed that Unigram is not the only way to measure relevance. For example, Fresneda & Gefen (2019) evaluated words "unicity" in terms of message entropy.

Ghose & Ipeirotis (2011) analyzed the **readability** of reviews and their **linguistic correctness** (lack of misspellings, etc.), both of which are observed to positively influence perceived helpfulness. However, Liu et al. (2019) proved that readability depends on how closely a review matches the language style of the target readers. Therefore it is a domain-dependent indicator.

In order to base our work on largely applicable helpfulness determinants, we focus our analysis on length, rating and Unigram, leaving linguistic correctness and readability apart. Moreover, we study deviations in length and rating by grouping reviews by author or item in order to provide a contextual analysis of user feedback. Finally, we test the usefulness of our helpfulness estimation model for decision-making by integrating it into a collaborative recommender system and by measuring the improvements in suggestion performance.

2.2.2. Semantic features

Cao et al. (2011) and Qazi et al. (2016) found that **semantic features** of reviews positively influence helpfulness perception. Indeed, the semantic

analysis includes diverse approaches that also exploit some structural features, such as the number of product attributes mentioned in a review, and the length of its sentences (Sun et al., 2019).

Among the identified helpfulness predictors there are the **positive or negative sentiment (polarity)** of reviews, combined with the number of positive/negative words (Dong et al., 2013). Eslami et al. (2018) observed that the most helpful comments are associated to medium length, lower scores, and negative or neutral polarity. Ahmad & Laroche (2017) noticed that negative reviews containing service failure data and positive reviews describing core product functionalities, technical aspects and aesthetics are perceived as helpful. Salehan & Kim (2016) found that sentimental reviews with neutral polarity in their text are perceived to be more helpful than the other ones.

Differently, Ghose & Ipeirotis (2011) discovered that very **objective** and very **subjective** comments are considered as helpful but mixed comments are not. Moreover, Krishnamoorthy (2015) observed that syntactic structure and presence of adjectives, state and action verbs are good helpfulness predictors, especially if used in conjunction with readability and subjectivity, review age and rating.

Aspect-based approaches for helpfulness assessment employ techniques such as Supervised LDA (Blei & McAuliffe, 2007) and double propagation to extract aspects from reviews as latent topics. See Xiong & Litman (2014) and Paul et al. (2017), respectively. However, Yang et al. (2016) noticed that LDA produces a large number of low-level, product-dependent aspects.

We aim at developing a model that can be transferred to different service domains. For this purpose, we focus on review polarity, which we analyze both in absolute terms, as done in previous work, and contextually, from the viewpoint of user/item-based deviations. Moreover, we integrate our model into collaborative recommendation.

2.2.3. Consistency and Rating Deviations

Some recommender systems use **consistency** (henceforth, **coherence**) to evaluate reviewers' reliability, having observed that large discrepancies between review sentiment and rating can be a sign of low-quality (Shen et al., 2019). Also (Dong et al., 2013) investigated this feature but they have not described its impact on perceived helpfulness. Zhou et al. (2020) have looked at consistency from a different perspective and they discovered that the similarity between review title and review content positively influences perceived helpfulness.

Raghavan et al. (2012) found that review length and the **deviation of the rating** from the mean rating of the product are strong helpfulness predictors. Moreover, they observed that the regression models that use these features perform better than those relying on semantic features, either based on TF/IDF or LDA.

In our helpfulness estimation model we include the rating-polarity coherence as a candidate determinant but we omit the analysis of the similarity between title and content because, as described in Section 3.1, the reviews used for our experiments have no title. However, our model could be seamlessly extended to consider this additional element.

2.3. Moderating factors

Two *moderating factors* of helpfulness perception can influence readers' voting behavior:

- The first is **Product type**. Mudambi & Schuff (2010) observed that extreme ratings negatively influence perceived helpfulness in *experience goods*, while review length has greater positive effect on *search goods* than on experience ones. Moreover, Siering et al. (2018) found that the strength of sentiment increases review helpfulness for search products while it decreases helpfulness for experience products.[1]

[1] According to Nelson (1974), *search goods* are products for which the consumer can obtain information about quality prior to purchase. Differently, *experience goods* require sampling or

- The second factor is the **operationalization of perceived helpfulness**; in other words, its implementation. Wu (2017) analyzed Amazon.com (2020) experience products and found that, considering the ratio between the number of positive votes and the total number of votes, review valence (intended as rating) positively influences helpfulness. However, the opposite result is obtained if helpfulness is computed as the count of votes received by a review. Moreover, Hong et al. (2017) discovered that, while review length is a significant determinant of helpfulness, regardless of its operationalization, it has stronger effect when it is measured as the count of votes. More generally, Hong et al. (2017) found that review length, review age and reviewer expertise positively influence perceived helpfulness while readability and rating are insignificant determinants, regardless of the applied helpfulness measure.

In our experiments we focus on accommodation and food services, all of which are classified as search products. Therefore, our analysis is not particularly affected by the effect of product moderation. Moreover, we operationalize perceived helpfulness as the count of votes because we use review datasets which include positive feedback about reviews.

2.4. Summary of our work

We focus on review-related determinants and we leave apart reviewer-related properties (expertise, reputation, productivity, anonymity, trustworthiness, etc. (Malik & Hussain, 2018; Filieri et al., 2018; Siering et al., 2018; Davis & Agrawal, 2018)) and context (review age, visibility, etc., (Hu & Chen, 2016; Hu et al., 2017)). We exclude these aspects in order to restrict the number of factors to be analyzed. We also exclude the analysis of the hedonic value of reviews (Ham et al., 2019) because we focus on decision-making-related aspects.

While some works have studied the deviation between the rating of a review and the mean rating of the item involved, our work introduces the deviations

purchase to evaluate their quality.

with respect to length, polarity and coherence, by user and by item, in order to understand whether this is helpful information to item evaluation. Moreover, we provide a prediction model that we validate by means of correlation analysis using observed perceived helpfulness, and by applying the model to a collaborative recommendation algorithm.

3. Research Methodology

In order to focus on a set of largely-recognized helpfulness determinants, we select length, Unigram, rating, polarity and coherence as basic factors to be investigated and we study the deviations in the values of these factors from average, user-based or item-based. Specifically, we abstract from domain-dependent semantic concepts, which lack generalizability, and we only exploit Unigram as a lightweight measure for the assessment of the amount of content provided by reviews. We also exclude review age because it is a partial indicator, as the datasets we use provide no information about the visibility of reviews (Hu & Chen, 2016).

We analyse dependencies between factors and perceived helpfulness by exploiting regression models to understand the influence of determinants. Other works employ neural networks in learning data to use it for prediction models. See Fan et al. (2019) and Malik & Hussain (2017). However, those approaches fail to shed light on the impact of individual features. In other words, they discover which combination of factors achieves the best results but they cannot reveal the influence of individual determinants on review helpfulness. Moreover, comparing a regression-based approach (Yang et al., 2016) with an advanced neural one (Chen et al., 2019) on the same dataset shows that the regression model performs almost as well as the neural one. While this might not be true in general, we prefer regression because of the transparency of its results.

Below, we introduce notation used in the following sections:

- $\mathcal{I} = \{i_1, \ldots, i_m\}$ is the set of items (products or services);

Table 2: Descriptive statistics of variables.

	Yelp-Hotel						Yelp-Food					
	Count	Min	Max	Mean	STD	Median	Count	Min	Max	Mean	STD	Median
Number of reviews (+ ratings)	10081						65120					
Number of users	654						3105					
Number of items	1081						2150					
Number of reviews x user		10	106	15.4144	9.1463	13		10	320	20.9726	18.2355	15
Number of reviews x item		1	222	9.3256	27.1735	2		1	485	30.2884	49.4937	12
Number of helpfulness votes x review		0	559	7.3118	18.1625	3		0	227	4.5535	9.1838	2
Rating values		1	5	3.5604	1.0661	4		1	5	3.7904	1.1519	4
Review length		4	1005	175.5369	145.6903	134		1	1011	137.2749	117.2275	104
Review polarity		1.1047	4.9310	3.9499	0.5796	4.1446		1.0989	4.9681	4.0086	0.5583	4.1774
Rating-polarity coherence		1.6385	5.0000	4.2291	0.6276	4.3157		1.1721	5	4.2388	0.6135	4.3364
STD of rating values x user		0	1.9315	0.9614				0	2.0248	1.0743		
STD of rating values x item		0	2.8284	0.7928				0	2.8284	1.0827		
STD of review polarity x user		0.0485	1.2782	0.4741				0.0444	1.3151	0.4977		
STD of review polarity x item		0.0002	1.6106	0.3931				0.0006	1.7474	0.5441		
STD of review length x user		12.0504	303.2880	94.3964				3.4667	325.8014	72.2702		
STD of review length x item		0.7071	536.6940	117.1095				0	627.9108	99.2652		

- $\mathcal{U} = \{u_1, \ldots, u_n\}$ is the set of users - users can post reviews about items and vote the helpfulness of the reviews written by the other people;
- $\mathcal{R} = \{r_1, \ldots, r_k\}$ is the set of reviews. We assume that each comment is associated with a rating of the reviewed item.

3.1. Data

For our analysis we use two subsets of the (Yelp, 2019b) dataset:

- YELP-Hotel stores reviews about accommodation services;
- YELP-Food stores reviews about food services in the city of Phoenix.

In (Yelp, 2019b), each item (business) is associated with a list of tags representing the categories to which it belongs. We obtained these datasets from the main one by applying two filters. First, we filtered items by tag and we removed the items which had no associated review+rating. Then, we removed the infor-

mation about the users who provided less than 10 reviews. This is important to support the analysis of deviations in individual user behavior.[2]

In the datasets, each business is associated with the rating scores and free text reviews provided by Yelp users. Item ratings take values in a [1,5] Likert scale where 1 is the worst value and 5 is the best one. Moreover each review is associated with the feedback it receives from its own readers (for example, "useful" votes). Yelp only supports the expression of positive feedback. Table 2 provides some descriptive statistics about the two datasets:

- The higher portion of the table reports general statistics about the dataset ("Number of reviews (+ ratings)", ..., "Number of helpfulness votes × review"). Looking at line "Number of reviews × user", which shows how many comments have been provided by individual users, we notice that this distribution has a long tail. Few users wrote many reviews; the other people provided very few ones. The distributions of the number of reviews × item and that of helpfulness votes × review are similar.

- The second portion of the table ("Rating values", ..., "Rating-Polarity coherence") summarizes the distribution of rating scores on items, and statistics regarding review length and polarity. We compute the polarity of each comment as follows:

  - First, we retrieve the polarity values generated by TextBlob (Loria, 2020) and VADER (Hutto & Eric, 2014);
  - Then, we compute the mean value among the two and we convert it in the [1, 5] interval.

  In this way, the final polarity value can be compared with the rating associated to the review for the evaluation of the Rating-polarity coherence.

[2]The full list of Yelp categories is available at `https://www.yelp.com/developers/documentation/v3/category_list`. Appendix A reports the categories we used to produce the two datasets.

We notice that reviews are fairly consistent, with a value of 4.229 in the [1, 5] interval.

- The third portion of the table ("STD of rating values × user", ..., "STD of review length × item") provides information about the standard deviation of rating scores, polarity and length across users or items. Rating and reviewing behavior is not uniform. Specifically, we observe differences in the expression of ratings (mean STD=0.96 in [1, 5]), polarity (mean STD=0.474) and length (mean STD = 94 words). Analogous considerations can be made looking at reviews from the viewpoint of items, even though, in that case, the standard deviation is a bit lower.

The observations concerning standard deviations in reviews written by the same users, or concerning the same items, highlight the relevance of studying these aspects. In the following we describe the dependent and independent variables we defined and the analysis method we applied.

3.2. Research variables

3.2.1. Dependent variable

Given a review $r \in \mathcal{R}$ concerning an item $i \in \mathcal{I}$, the only dependent variable we consider is the **perceived helpfulness** of r, which we operationalize in terms of counting votes, normalized in the [0, 1] interval:

$$PerceivedHelpfulness_r = f(|Votes_r|) \tag{1}$$

In Equation 1, $Votes_r$ is the total number of votes received by r. This is the sum of "useful", "funny" and "cool" votes given to r. Moreover $|.|$ is set cardinality. Function $f()$ normalizes its argument in the [0, 1] interval:

$$f(x) = \frac{log(x+1)}{1+log(x+1)} \tag{2}$$

We adopt this operationalization because, as previously specified, Yelp only supports positive feedback.

Table 3: Independent variables and their operationalization: $r \in \mathcal{R}$ denotes a review about an item $i \in \mathcal{I}$ and $u \in \mathcal{U}$ is the author of r. $Words_r$ is the set of words included in r. $Revs_u$ is the set of reviews written by u. $Revs_i$ is the set of reviews about i and $polarity_r$ is the polarity of r computed as explained in Section 3.1. All the variables are normalized in [0,1] using formula $f()$ of Equation 2.

Variable	Operationalization	Notes
RAT_r	$RAT_r = f(rating_r)$	$rating_r$ is the rating included in r
LEN_r	$LEN_r = f(\|Words_r\|)$	$\|.\|$ denotes set cardinality
UGR_r	$UGR_r = f(TF\_IDF)$	$TF\_IDF_r = \frac{\sum_{w \in Words_r} TF\_IDF_w}{\|Words_r\|}$
POL_r	$f(polarity_r)$	
COH_r	$f(1 - \|\|RAT_r - POL_r\|\|)$	$\|\|.\|\|$ denotes absolute value
ΔLEN_{ru}	$f(\|\|LEN_r - \frac{\sum_{x \in Revs_u} LEN_x}{\|Revs_u\|}\|\|)$	
ΔLEN_{ri}	$f(\|\|LEN_r - \frac{\sum_{x \in Revs_i} LEN_x}{\|Revs_i\|}\|\|)$	
ΔRAT_{ru}	$f(\|\|RAT_r - \frac{\sum_{x \in Revs_u} RAT_x}{\|Revs_u\|}\|\|)$	
ΔRAT_{ri}	$f(\|\|RAT_r - \frac{\sum_{x \in Revs_i} RAT_x}{\|Revs_i\|}\|\|)$	
ΔPOL_{ru}	$f(\|\|POL_r - \frac{\sum_{x \in Revs_u} POL_x}{\|Revs_u\|}\|\|)$	
ΔPOL_{ri}	$f(\|\|POL_r - \frac{\sum_{x \in Revs_i} POL_x}{\|Revs_i\|}\|\|)$	

3.2.2. Independent variables

Given a review $r \in \mathcal{R}$ about an item $i \in \mathcal{I}$, we consider the following independent variables. Table 3 shows the operationalizations of these variables, all of which are normalized in [0, 1] using Equation 2:

- RAT_r**:** normalized rating of i in r. This is the normalized score value that r's author attributed to item i;
- LEN_r**:** normalized number of words included in r;
- UGR_r *(Unigram)*: normalized, mean TF/IDF value (Robertson, 2004) of the lemmatized words included in r after having removed stop words and very short words (composed of maximum 2 letters) from the text;
- POL_r**:** normalized polarity of the text of r, computed as explained in Section 3.1;
- COH_r**:** normalized coherence between the polarity of r and the rating of item i;
- ΔLEN_{ru}**:** normalized absolute distance between the length of r and the mean length of the reviews written by u;
- ΔLEN_{ri}**:** normalized absolute distance between the length of r and the mean length of the reviews about item i;
- ΔRAT_{ru}**:** normalized absolute distance from the rating of i in r and the mean rating of items in the reviews written by u;
- ΔRAT_{ri}**:** normalized absolute distance between the rating of i in r and the mean rating of i;
- ΔPOL_{ru}**:** normalized absolute distance between the polarity of r and the mean polarity of the reviews written by u;
- ΔPOL_{ri}**:** normalized absolute distance between the polarity of r and the mean polarity of the reviews about i.

Table 4: Pearson correlation coefficients between variables on Yelp-Hotel dataset.

		1	2	3	4	5	6	7	8	9	10	11	12
1	RAT_r	1											
2	LEN_r	-0.0276	1										
3	UGR_r	0.0109	-0.9614	1									
4	POL_r	0.5297	0.0525	-0.0644	1								
5	COH_r	0.6446	-0.0436	0.0320	0.0856	1							
6	ΔLEN_{ru}	-0.0487	0.1592	-0.1340	-0.0338	-0.0385	1						
7	ΔLEN_{ri}	-0.0018	-0.0529	0.0761	-0.0158	-0.0046	0.0566	1					
8	ΔRAT_{ru}	-0.3257	0.0197	-0.0156	-0.2151	-0.4185	0.0494	0.0265	1				
9	ΔRAT_{ri}	-0.2840	-0.0272	0.0408	-0.2113	-0.3279	0.0162	0.3447	0.4061	1			
10	ΔPOL_{ru}	-0.2994	-0.1782	0.1791	-0.6683	-0.0078	-0.0015	0.0449	0.2616	0.1942	1		
11	ΔPOL_{ri}	-0.3158	-0.1406	0.1510	-0.6649	-0.0373	-0.0034	0.2582	0.1456	0.3306	0.6383	1	
12	$PerceivedHelpfulness_r$	-0.0531	0.3619	-0.3567	-0.0133	-0.0625	0.1140	-0.0026	0.0707	0.0415	-0.0696	-0.0428	1

Table 5: Pearson correlation coefficients between variables on Yelp-Food dataset.

		1	2	3	4	5	6	7	8	9	10	11	12
1	RAT_r	1											
2	LEN_r	-0.1058	1										
3	UGR_r	0.0728	-0.9594	1									
4	POL_r	0.5879	0.0368	-0.0614	1								
5	COH_r	0.6323	-0.0891	0.0679	0.1385	1							
6	ΔLEN_{ru}	-0.0616	0.1506	-0.1425	-0.0166	-0.0659	1						
7	ΔLEN_{ri}	0.0117	-0.0825	0.0890	0.0249	-0.0162	0.1076	1					
8	ΔRAT_{ru}	-0.4164	0.0206	-0.0061	-0.3053	-0.5346	0.0490	0.0117	1				
9	ΔRAT_{ri}	-0.3944	-0.0056	0.0200	-0.2846	-0.5443	0.0180	0.0553	0.5992	1			
10	ΔPOL_{ru}	-0.3517	-0.2063	0.2173	-0.6771	-0.0428	-0.0205	0.0050	0.3298	0.2239	1		
11	ΔPOL_{ri}	-0.3571	-0.2052	0.2218	-0.6837	-0.0313	-0.0292	0.0243	0.2100	0.2915	0.7193	1	
12	$PerceivedHelpfulness_r$	-0.0720	0.3685	-0.3702	-0.0319	-0.0719	0.1467	-0.0026	0.0510	0.0319	-0.0663	-0.0604	1

3.2.3. Correlation analysis

Table 4 shows the Pearson correlation coefficient between each couple of independent variables and between the variables and the observed helpfulness ($PerceivedHelpfulness_r$) in the Yelp-Hotel dataset. Table 5 provides the same type of information for the Yelp-Food dataset. In the following, we jointly discuss the two sets of results because we observe similar correlation results.

Review polarity (POL_r) is highly correlated with the associated rating RAT_r ($r_{Pearson} = 0.5297$ in Yelp-Hotel, $r_{Pearson} = 0.5879$ in Yelp-Food). This finding is consistent with the good "Rating-polarity coherence" observed in Table 2 and suggests that the rating is mostly in line with the valence of review text. Moreover ΔRAT_{ri} is highly correlated with ΔRAT_{ru} ($r_{Pearson} = 0.4061$ in

Yelp-Hotel, $r_{Pearson} = 0.6992$ in Yelp-Food). Thus, the difference between the rating of an item i and its average rating, and the difference between the rating of i and the mean ratings provided by the same user, are similar. The tables also show that there is a correlation between ΔPOL_{ri} and ΔPOL_{ru}, denoting that a similar behavior is observed for review valence. Finally, $PerceivedHelpfulness_r$ correlates well with review length (LEN_r).

As described in Section 3.3, in order to understand how variables interact with each other, and whether we can ignore some of them in perceived helpfulness estimation, we perform a regression analysis on the models that are based on these variables.

3.3. Empirical model

We consider two baseline helpfulness estimation models that include traditional determinants, and our proposed one:

- M1: $\beta_0 + \beta_1 RAT_r + \beta_2 LEN_r + \beta_3 UGR_r$;
- M2: $\beta_0 + \beta_1 RAT_r + \beta_2 LEN_r + \beta_3 UGR_r + \beta_4 POL_r$;
- M3: $\beta_0 + \beta_1 RAT_r + \beta_2 LEN_r + \beta_3 UGR_r + \beta_4 POL_r + \beta_5 COH_r + \beta_6 \Delta LEN_{ru} + \beta_7 \Delta LEN_{ri} + \beta_8 \Delta RAT_{ru} + \beta_9 \Delta RAT_{ri} + \beta_{10} \Delta POL_{ru} + \beta_{11} \Delta POL_{ri}$.

We learn two versions of each model Mj ($j \in \{1, 2, 3\}$):

1. Mj_L is obtained by means of linear regression. For this version, we use Linear Support Vector Regression implemented in the `scikit-learn` library (Pedregosa et al., 2011), which supports the analysis of the positive or negative influence of factors on helpfulness perception.
2. Mj_{NL} is obtained by means of a regression algorithm which can identify non linear dependencies. For this version, we use Random Forest Regression implemented in the `scikit-learn` library (Pedregosa et al., 2011).

We evaluate the models by comparing the estimated helpfulness values they generate with the helpfulness observed in the datasets by means of Pearson and Spearman correlation analyses.

Table 6: Pearson and Spearman correlation values of the M1, M2 and M3 models learned on Yelp-Hotel and Yelp-Food using Linear Support Vector Regression and Random Forest Regression. The best results are in bold. Significance is encoded as (**) $p < 0.01$.

	Yelp-Hotel		Yelp-Food	
	Pearson's r	Spearman's r	Pearson's r	Spearman's r
$M1_L$	0.3459**	0.3652**	0.3827**	0.4114**
$M2_L$	0.3460**	0.3644**	0.3849**	0.4137**
$M3_L$	0.3635**	0.3725**	0.3982**	0.4169**
$M1_{NL}$	0.2952**	0.3002**	0.3185**	0.3240**
$M2_{NL}$	0.3065**	0.3106**	0.3675**	0.3729**
$M3_{NL}$	**0.4071****	**0.4036****	**0.4418****	**0.4486****

4. Results

4.1. Performance of the helpfulness estimation models

Table 6 show the results of Linear Support Vector Regression and Random Forest Regression applied to M1, M2 and M3 on Yelp-Hotel and Yelp-Food. For each dataset we carried out the experiments by applying a 5-fold cross validation in order to avoid biases related to the way the dataset is split. We can notice that M3 outperforms M1 and M2 in terms of Pearson and Spearman correlation both when learned through linear ($M3_L$) and non linear ($M3_{NL}$) regression. This means that, by adding the deviations of rating, polarity and length to the variables used in M_1 and M_2, we improve helpfulness prediction. Overall, $M3_{NL}$ obtains the best results. In other words, the Random Forest Regressor is a better helpfulness predictor than Linear Support Vector Regression on the datasets we analyzed.

Table 7: Analysis on the two datasets by means of Linear Support Vector Regression. For each model, the coefficients show the impact of variables on perceived helpfulness. Significance is encoded as (**) $p < 0.01$ and (*) $p < 0.05$.

Variable	Yelp-Hotel			Yelp-Food		
	$M1_L$	$M2_L$	$M3_L$	$M1_L$	$M2_L$	$M3_L$
RAT_r	-0.1905**	-0.1560**		-0.1662**	-0.0741**	0.0963**
LEN_r	1.5740**	1.5836**		0.7060**	0.7016**	0.6275**
UGR_r	-0.8825	-0.8820	-4.1487**	-2.5234**	-2.5789**	-2.4575**
POL_r		-0.1330	-0.5983**		-0.3712**	-0.7557**
COH_r			-0.1008			-0.2534**
ΔLEN_{ru}			0.1395**			0.1814**
ΔLEN_{ri}			0.0222			
ΔRAT_{ru}			0.1041**			0.0520**
ΔRAT_{ri}			0.0571*			
ΔPOL_{ru}			-0.1500**			-0.1231**
ΔPOL_{ri}			-0.0455			

4.2. Impact of independent variables on review helpfulness

4.2.1. Linear SVR regression results

Table 7 shows the weights assigned to the variables by the Linear Support Vector Regression on Yelp-Hotel and Yelp-Food. Most coefficients are statistically significant. First, we analyze the perceived helpfulness determinants identified in the previous research. Then we study the ones we propose.

On both datasets, UGR_r negatively influences perceived helpfulness in all

the models and the same happens for POL_r in $M2_L$ and $M3_L$. Furthermore RAT_r has a negative influence, with the exception of $M3_L$ where this factor is not used, or it has a slightly positive effect. Differently, LEN_r has a positive impact on helpfulness in all the models, except for $M3_L$ that does not use it on Yelp-Hotel. These results are fairly consistent with the previous research results, some of which have investigated the importance of these factors, but not the positive nor negative direction of influence.

As far as models $M2_L$ and $M3_L$ are concerned, we observe that COH_r is used on both datasets with a negative weight. While this is apparently counter-intuitive, it must be noted that the consistency between review sentiment and rating does not mean that the review is helpful. In fact, in the recommender systems research, this type of information is used to assess reviewers' reliability, which is different from review helpfulness (Shen et al., 2019).

The results concerning the deviations on length, rating and polarity are not completely aligned but they are fairly consistent:

- On Yelp-Hotel the regression model uses all these factors for prediction. While the deviations in polarity have negative impact on perceived helpfulness, the deviations in length and rating positively influence this variable. We also observe that the user-based deviations have stronger influence than item-based ones, whose impact is low;

- On Yelp-Food the impact of user-based length and rating deviations are consistent with those of the other dataset, while item-based deviations are not used.

4.2.2. Random Forest Regression results

Table 8 shows the importance of the independent variables using Random Forest Regression. UGR_r, which represents the content of the reviews, strongly affects helpfulness prediction in all the models and datasets. It is the most influential determinant among the traditional ones (rating, length and Unigram).

Table 8: Analysis on the two datasets using Random Forest Regression. For each model, the values show the importance of the corresponding variables.

Variable	Yelp-Hotel			Yelp-Food		
	$M1_{NL}$	$M2_{NL}$	$M3_{NL}$	$M1_{NL}$	$M2_{NL}$	$M3_{NL}$
RAT_r	0.0352	0.0637	0.0346	0.0187	0.0435	0.0208
LEN_r	0.4129	0.3023	0.1949	0.3206	0.2571	0.1834
UGR_r	0.5520	0.3574	0.2257	0.6608	0.4130	0.2400
POL_r		0.2766			0.2864	
COH_r						0.1313
ΔLEN_{ru}			0.1559			0.1602
ΔLEN_{ri}						
ΔRAT_{ru}			0.1378			0.1299
ΔRAT_{ri}			0.1202			
ΔPOL_{ru}						0.1343
ΔPOL_{ri}			0.1309			

LEN_r is quite influential as well: it has the second highest importance in all the models. RAT_r is used everywhere but with minor importance.

Interestingly, POL_r is rather influent in $M2_{NL}$ on both datasets. However, it disappears in $M3_{NL}$, which exploits user-based and/or item-based deviations. This suggests that these factors are more predictive than polarity. Moreover, COH_r is only used on Yelp-Food.

In $M3_{NL}$ the Random Forest Regressor uses different deviation variables for the prediction:

- ΔLEN_{ru} and ΔRAT_{ru} are used in both datasets;
- ΔPOL_{ru} is only recognized as influential in Yelp-Food;
- ΔRAT_{ri} is only used in Yelp-Hotel;
- ΔLEN_{ri} is ignored in both datasets;

- ΔPOL_{ru} is only used in Yelp-Food;
- ΔPOL_{ri} is only used in Yelp-Hotel.

This shows that, while some variables, such as rating, length and Unigram are influential, the impact of coherence and item-based ratings depends on product type. In particular, for Yelp-Food user-based deviations are meaningful while item-based ones are not. In contrast, the impact of user-based deviations, especially concerning review length and rating, is consistent across datasets.

Overall, we notice that variables representing deviations are less important than ratings, length and Unigram. However they clearly help improving perceived helpfulness prediction, as shown in Table 6.

5. Helpfulness-aware, personalized item recommendation

As described by Ricci et al. (2011), recommender systems leverage data about users' past rating behavior to personalize the suggestion of items. However, they uniformly exploit this type of information to predict ratings, without considering its quality. We point out that the usage of poor data might decrease recommendation performance because item evaluation would be based on unreliable information. With this perspective, we can further check the efficacy of our helpful estimation model by comparing a state of the art recommender system with an algorithm that tunes the impact of ratings in item evaluation on the basis of the predicted helpfulness of reviews. We develop our SVD$_{Helpfulness}$ recommender system by modifying SVD (Koren & Bell, 2011), which is a largely-used collaborative recommender system based on Matrix Factorization. See Sections 5.1 and 5.2. Then, we compare the accuracy, error minimization and ranking capabilities of SVD$_{Helpfulness}$ with those of SVD++ (Koren, 2008), which is a well-established baseline to evaluate recommender systems. See Section 5.3. For the development of SVD$_{Helpfulness}$ we predict the helpfulness of reviews by means of a hybrid model. Given a review $r \in \mathcal{R}$,

$PredictedHelpfulness_r$ is computed as follows:

$$PredictedHelpfulness_r = \begin{cases} PerceivedHelpfulness_r & \text{if } |Votes_r| > 0 \\ \text{value estimated by M3}_{NL} & \text{otherwise} \end{cases} \quad (3)$$

If the ground-truth helpfulness is available we use it. Otherwise, $PredictedHelpfulness_r$ is the value obtained by applying M3_{NL}, which is the best performing model according to the analysis of Section 4. Notice that we tested other combinations of the two measures, including their average, but the final performance of the recommender was lower. Therefore we selected this approach for our experiments. Overall, we use the helpfulness values estimated by M3_{NL} in 21% of Yelp-Hotel reviews, and in 30% of Yelp-Food reviews.

5.1. Collaborative Filtering with Matrix Factorization

The recommender systems based on Matrix Factorization assume that a few latent patterns influence rating behavior. These systems perform a low-rank matrix factorization on the users-items rating matrix, which stores the evaluations of items provided by users (Koren & Bell, 2011). We assume that there are n users and m items and we adopt the following notation:

- $\mathbf{R} \in \mathbb{R}^{n \times m}$ is the users-items rating matrix;
- $\mathbf{R}_{xy}$ is the rating given by user $u_x \in \mathcal{U}$ to item $i_y \in \mathcal{I}$, if any:
  - $\mathcal{O} = \{< u_x, i_y > \mid \mathbf{R}_{xy} \neq 0\}$ is the set of observed ratings. This set includes all the $< u_x, i_y >$ pairs such that, as reported in the dataset, user u_x has given a rating in [1, 5] to item i_y.
  - $\mathcal{T} = \{< u_x, i_y > \mid \mathbf{R}_{xy} = 0\}$ is the set of unknown ratings.

We assume that there are K latent factors; then:

- $\mathbf{u}_x \in \mathbb{R}^K$ denotes the user preference vector of user u_x and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{K \times n}$ stores the preference vectors of all the users;
- $\mathbf{i}_y \in \mathbb{R}^K$ denotes the item characteristic vector of i_y and $\mathbf{I} = [\mathbf{i}_1, \ldots, \mathbf{i}_m] \in \mathbb{R}^{K \times m}$ stores the item characteristic vectors of all the items.

In order to learn these vectors, the recommender system solves the following optimization problem:

$$\min_{\mathbf{U},\mathbf{I}} \sum_{<u_x,i_y>\in\mathcal{O}} (\mathbf{R}_{xy} - \mathbf{u}_x^T \mathbf{i}_y)^2 + \lambda(||\mathbf{U}||_F^2 + ||\mathbf{I}||_F^2) \tag{4}$$

Equation 4 is aimed at finding a setting of $\mathbf{U}$ and $\mathbf{I}$ that minimizes the distance between the observed ratings ($\mathcal{O}$) and the estimated ones (obtained as the product of transposed user preference vectors and item characteristic ones). In the equation, $||.||_F$ denotes the Frobenius Norm and $||\mathbf{U}||_F^2 + ||\mathbf{I}||_F^2$ are the regularization terms to avoid over-fitting. Moreover $\lambda > 0$ controls the impact of $\mathbf{U}$ and $\mathbf{I}$ on regularization. The smaller is $\lambda > 0$, the minor is the influence of these vectors.

5.2. Steering Matrix Factorization by means of review helpfulness

In order to steer Matrix Factorization by means of review helpfulness, we introduce a weighting factor that tunes the impact of ratings depending on the predicted helpfulness of the associated reviews. The idea is that the ratings associated to highly helpful reviews should influence Matrix Factorization more than the other ones. The optimization problem is thus:

$$\min_{\mathbf{U},\mathbf{I}} \sum_{<u_x,i_y>\in\mathcal{O}} w_{xy}(\mathbf{R}_{xy} - \mathbf{u}_x^T \mathbf{i}_y)^2 + \lambda(||\mathbf{U}||_F^2 + ||\mathbf{I}||_F^2) \tag{5}$$

where w_{xy} is the predicted helpfulness of the review r associated to rating $\mathbf{R}_{xy}$, corresponding to $PredictedHelpfulness_r$ in Equation 3.

5.3. Validation

In this section we present the evaluation results of $\text{SVD}_{Helpfulness}$ compared with SVD++, using standard performance metrics for recommender systems (Jannach et al., 2016). We focus on ranking capability (MAP, MRR and NDCG), which is very important in the evaluation of recommender systems because it tells us how good an algorithm is at placing relevant items in the first positions of the suggestion list. Moreover we consider accuracy (Precision, Recall and F1) and error minimization (MAE and RMSE) metrics. For

Table 9: Recommendation performance of $\text{SVD}_{Helpfulness}$ and SVD++ on "Yelp-Hotel" and "Yelp-Food" datasets. Stars indicate significant differences according to a Wilcoxon signed-rank test between the two algorithms ((*) p <0.05). The percentages in brackets denote the relative difference between the values obtained by the two algorithms.

	Yelp-Hotel		Yelp-Food	
	$\text{SVD}_{Helpfulness}$	SVD++	$\text{SVD}_{Helpfulness}$	SVD++
Precision	0.7971* (+4.66%)	0.7616	0.7804* (+2.46%)	0.7617
Recall	0.7458* (+3.03%)	0.7239	0.7956* (+5.55%)	0.7538
F1	0.7706* (+3.81%)	0.7423	0.7879* (+3.99%)	0.7577
MAP	0.7159* (+3.83%)	0.6895	0.7385* (+6.60%)	0.6928
MRR	0.6313 (+1.61%)	0.6213	0.7678* (+1.59%)	0.7558
NDCG	0.9782* (+0.56%)	0.9728	0.9666* (+0.62%)	0.9606
RMSE	0.9279* (-6.13%)	0.9885	1.0655* (-7.39%)	1.1505
MAE	0.7206* (-0.73%)	0.7726	0.8314* (-7.47%)	0.8985

the evaluation, we perform a 5-fold cross validation, having set the number of latent factors to 50 and the learning rate to 0.01, and we optimize the other parameters by taking the configuration that achieves the best MAP as optimal.

Table 9 shows the results of the two algorithms on Yelp-Hotel and Yelp-Food and reports the relative differences in performance as percentages. Notice that, different from the other metrics, RMSE and MAE describe the error in rating estimation and thus have to be minimized. Therefore, the negative values associated to these measure denote an improvement in performance. On both datasets, $\text{SVD}_{Helpfulness}$ outperforms SVD++ in all the measures.

We can provide a more general view of performance by grouping measures according to high-level dimensions and by computing the mean values of the relative differences between algorithms. $\text{SVD}_{Helpfulness}$ compares to SVD++ as follows:

- The mean relative improvement of accuracy (Precision, Recall and F1) is equal to 3.83% on YELP-Hotel and by 3.99% on YELP-Food;

- The mean relative improvement of ranking capability (MAP, MRR and NDCG) is equal to 1.99% on Yelp-Hotel and 2.94% on Yelp-Food;
- Finally, $\text{SVD}_{Helpfulness}$ outperforms SVD++ in error minimization (RMSE and MAE) by 6.43% on YELP-Hotel and 7.43% on Yelp-Food.

6. Discussion

We investigated the impact on perceived review helpfulness of deviations in length, polarity and rating with respect to the reviews written by the same user, or concerning the same item. The results of our study show that the deviations from typical user behavior influence perceived helpfulness and support the identification of high-quality ratings in collaborative recommendation. By considering the results obtained using Random Forest Regression (which outperforms Linear Support Vector Regression in helpfulness estimation), we can answer our research questions as follows:

- *RQ1: Given a review r, does a deviation from the mean length, polarity and rating of the other reviews written by the same person provide useful information to assess the perceived helpfulness of r?*

  User-based deviations are useful to predict perceived review helpfulness. Specifically, the deviations concerning length and ratings influence helpfulness across both datasets that we considered, while the deviations in polarity have a more limited impact on it. These findings suggest that user-based deviations are important determinants to be considered within a review helpfulness estimation model. Particular attention should be given to the deviations in length and ratings which are, at the same time, stronger predictors and lighter measures to be computed than polarity. However, polarity should not be disregarded as a proxy of ratings in datasets that do not provide this type of information, such as the Airbnb (2020) one available at `http://insideairbnb.com/get-the-data.html`.

- *RQ2: Given a review r, does a deviation from the mean length, polarity and rating of the other reviews on the same item provide useful information to assess the perceived helpfulness of r?*

  The situation of item-based deviations is more complex because the results across datasets are heterogeneous. Deviations in review length can be ignored because they have no impact on helpfulness estimation. Differently, deviations in rating and in polarity are only useful in the Yelp-Hotel dataset. Therefore, the value of these indicators must be assessed before applying them to analyze review helpfulness in a specific domain.

6.1. Theoretical implications

Overall, these results advance the state of the art in helpfulness estimation, which traditionally focused on local properties of reviews (Mudambi & Schuff (2010); Yang et al. (2015); Hong et al. (2017); Siering et al. (2018)), or on rating deviations with respect to the average evaluation of an item (Raghavan et al. (2012); Fang et al. (2016)). The novel aspect of our work is the analysis of consumer feedback on a broader user or item-related context, considering a larger set of determinants.

As discussed in Section 5.3, the validation carried out by integrating our helpfulness estimation model into collaborative recommendation shows that the $\text{SVD}_{Helpfulness}$ algorithm sensibly outperforms SVD++ (which uniformly applies ratings for item estimation) in accuracy, ranking capability and error minimization on both datasets we considered. The superior performance of $\text{SVD}_{Helpfulness}$ is obviously relevant to the generation of good recommendation lists. Moreover, it has important implications regarding algorithmic experience (Shin et al., 2020) because it supports an explanation of suggestions based on high-quality feedback provided by the people who have previously experienced items. Specifically, a helpfulness-based exploitation of ratings and of the associated reviews can be the basis for the generation of explanations based on reliable user experience, as suggested by Ghose & Ipeirotis (2011) and preliminarily investigated by Mauro et al. (2020a).

6.2. Practical implications

The findings of this study are expected to provide insights for retailers and platform developers regarding how to suggest helpful reviews to consumers, out of the plethora of available ones, in order to reduce information overload and to support decision-making. As discussed by Salehan & Kim (2016), older reviews tend to attract more readerships and many products have thousands of comments, most of which "never receive any attention from consumers because they are at the end of a long list." Our idea is thus that of recommending reviews that provide useful content to support item selection and purchase decisions. Some online review platforms, such as TripAdvisor (2017), provide filters to explicitly select comments by facets (language, geographic location, etc.) but they overlook the usefulness of review content. Other platforms, such as Amazon.com (2020) and Airbnb (2020), apply ranking strategies to promote the reviews which they consider as the most effective ones. For instance, they show helpful positive and negative comments first. Moreover, some researchers have proposed heuristics aimed at re-ranking reviews in order to balance their visibility and to address the "rich gets richer" dilemma (Wang et al., 2020). Our work advances the state of the art by identifying novel helpfulness determinants with the idea that "out of the core" messages can be relevant information sources for item selection. This criterion could be combined with the existing ones in order to promote valuable consumer feedback as soon as it is posted online, regardless of its own popularity.

It is worth mentioning that we proposed a model aimed at identifying the most informative reviews to help consumers in decision-making. Another relevant perspective concerns retailers and service providers. In that case, consumer feedback can be analyzed to identify the positive and negative aspects of products and services observed by consumers, with the aim of highlighting aspects that can be improved or promoted. For instance, see Qi et al. (2016), Xu & Lu (2016), Bilici & Saygın (2017), Prado & Moro (2017) and Xu et al. (2017). We are carrying out preliminary investigations in this direction to evaluate the helpfulness determinants we proposed in that context.

7. Limitations and future work

Our work has limitations that we would like to address:

1. Helpfulness estimation also concerns the properties of reviewers like their expertise (Siering et al., 2018), reputation (Tang et al., 2013; Chua & Banerjee, 2015; Huang et al., 2015; Gao et al., 2017; Jiang & Diesner, 2016), engagement and social influence (Ngo-Ye & Sinha, 2014; Mohammadiani et al., 2017). Within this extensive scenario, we contribute to enhance the content-analysis perspective. However, we plan to extend our analysis to the other variables.
2. We tested our model on two datasets concerning food and accommodation services. Further experiments with different datasets are needed to assess the validity of the model in other domains, such as the sales of search and experience products.
3. We describe the semantic features of reviews by focusing on TF/IDF to measure the relative importance of the words occurring in the reviews. In our future work we plan to integrate in our model an analysis of the role of emotions in order to enrich the type of information used for helpfulness assessment (Martin & Pu, 2014; Yang et al., 2015; Malik & Hussain, 2017; Gang & Hong, 2019).

We also plan to extend our work regarding review-aware recommender systems (Chen et al., 2015; Hernández-Rubio et al., 2019). Previously, Mauro et al. (2019) and Ardissono & Mauro (2020) investigated reviewer behavior with the aim of estimating reputation and they leveraged this information in a trust-based recommender system. Now, we plan to extend that work with the analysis of review helpfulness which, in turn, affects reviewers' trustworthiness.

Another interesting research path is the enhancement of data interpretation via information visualization, which we investigated in our recent research about information exploration (Mauro et al., 2019, 2020b). This is relevant to reveal interesting behavior patterns related to the influence of context (for instance,

travel context in (Chang et al., 2019)) and cultural background (Nakayama & Wan, 2019) on the aspects and evaluations appearing in the reviews.

8. Conclusions

This paper presented a study on perceived review helpfulness aimed at advancing the state of the art in the identification of the reviews which provide useful information to consumers for decision-making. We proposed a novel perspective on review analysis by considering *deviations in rating, length and valence* with respect to the reviews written by the same user, or concerning the same item.

We studied this phenomenon by developing three models which leverage different sets of content features: from traditional ones (length, rating, Unigram and polarity) to the novel determinants we propose, including coherence between rating and polarity, and deviations in length, polarity and rating pivoted on users and on items. We learned these models through regression on two large datasets about accommodation and food services. Our experiments showed that the analysis of these factors enhances the estimation of review helpfulness by reaching higher correlation values with the helpfulness feedback observed in the datasets. Specifically, the experimental results show that user-based deviations, especially regarding review length and rating, influence perceived helpfulness, while item-based deviations are weaker predictors. A further experiment in which we integrated helpfulness estimation into a collaborative recommender system has shown that this type of information enhances suggestion performance with respect to only using item ratings. In other words, the reviews estimated as helpful provide high-quality content for recommendation.

Overall these results are encouraging and suggest that our model is an effective tool to select relevant user feedback for decision-making.

9. Acknowledgments

This work was supported by the University of Torino (grant number: ARDL_RILO_19_01) which funded the conduct of the research and preparation of the article. We are grateful to Ms. Jeanne Marie Griffin for having proofread our paper.

Appendix A.

- The selection of businesses to define the Yelp-Hotel dataset is based on the following tags: Hotels, Mountain Huts, Residences, Rest Stops, Bed & Breakfast, Hostels, Resorts.
- The selection for Yelp-Food is based on the following tags: American, Argentine, Asian Fusion, Australian, Austrian, Bangladeshi, Belgian, Brasseries, Brazilian, British, Cambodian, Cantonese, Catalan, Chinese, Conveyor Belt Sushi, Cuban, Czech, Delis, Empanadas, Falafel, Filipino, Fish & Chips, French, German, Greek, Hawaiian, Himalayan/Nepalese, Hot Pot, Hungarian, Iberian, Indian, Indonesian, Irish, Italian, Japanese, Japanese Curry, Korean, Latin American, Lebanese, Malaysian, Mediterranean, Mexican, Middle Eastern, Modern European, Mongolian, New Mexican Cuisine, Noodles, Pakistani, Pan Asian, Persian/Iranian, Peruvian, Piadina, Pizza, Poke, Polish, Polynesian, Portuguese, Ramen, Russian, Salad, Scandinavian, Scottish, Seafood, Shanghainese, Sicilian, Singaporean, Soup, Southern, Spanish, Sri Lankan, Steakhouses, Sushi Bars, Syrian, Tacos, Tapas Bars, Tapas/Small Plates, Teppanyaki, Tex-Mex, Thai, Turkish, Ukrainian, Vegan, Vegetarian, Vietnamese, Wraps.

References

Ahmad, S. N., & Laroche, M. (2017). Analyzing electronic word of mouth: A social commerce construct. *International Journal of Information Management*,

37, 202 – 213. URL: http://www.sciencedirect.com/science/article/pii/S026840121630490X. doi:10.1016/j.ijinfomgt.2016.08.004.

Airbnb (2020). Airbnb. https://airbnb.com.

Amazon.com (2020). Amazon.com: online shopping for electronics, apparel, etc. http://www.amazon.com.

Ardissono, L., & Mauro, N. (2020). A compositional model of multi-faceted trust for personalized item recommendation. *Expert Systems with Applications*, *140*, 112880. URL: http://www.sciencedirect.com/science/article/pii/S0957417419305901. doi:https://doi.org/10.1016/j.eswa.2019.112880.

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* IUI '17 (p. 287–300). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3025171.3025209. doi:10.1145/3025171.3025209.

Berkovsky, S., Taib, R., Hijikata, Y., Braslavsku, P., & Knijnenburg, B. (2018). A cross-cultural analysis of trust in recommender systems. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* UMAP '18 (p. 285–289). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3209219.3209251. doi:10.1145/3209219.3209251.

Bilici, E., & Saygın, Y. (2017). Why do people (not) like me?: Mining opinion influencing factors from reviews. *Expert Systems with Applications*, *68*, 185 – 195. URL: http://www.sciencedirect.com/science/article/pii/S0957417416305322. doi:10.1016/j.eswa.2016.10.001.

Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* NIPS'07 (p. 121–128). Red Hook, NY, USA: Curran As-

sociates Inc. URL: https://dl.acm.org/doi/10.5555/2981562.2981578. doi:10.5555/2981562.2981578.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, *50*, 511 – 521. URL: http://www.sciencedirect.com/science/article/pii/S0167923610001909. doi:10.1016/j.dss.2010.11.009.

Chang, Y.-C., Ku, C.-H., & Chen, C.-H. (2019). Social media analytics: extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *International Journal of Information Management*, *48*, 263 – 279. URL: http://www.sciencedirect.com/science/article/pii/S0268401217303389. doi:10.1016/j.ijinfomgt.2017.11.001.

Chen, C., Qiu, M., Yang, Y., Zhou, J., Huang, J., Li, X., & Bao, F. S. (2019). Multi-domain gated CNN for review helpfulness prediction. In *The World Wide Web Conference* WWW '19 (p. 2630–2636). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3308558.3313587. doi:10.1145/3308558.3313587.

Chen, L., Chen, G., & Wang, F. (2015). Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*, *25*, 99–154. URL: https://doi.org/10.1007/s11257-015-9155-5. doi:10.1007/s11257-015-9155-5.

Chua, A., & Banerjee, S. (2015). Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth. *Journal of the Association for Information Science and Technology*, *66*, 354–362. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23180. doi:10.1002/asi.23180.

Davis, J. M., & Agrawal, D. (2018). Understanding the role of interpersonal identification in online review evaluation: An information processing perspective. *International Journal of Information Management*, *38*,

140 – 149. URL: http://www.sciencedirect.com/science/article/pii/S0268401216309057. doi:10.1016/j.ijinfomgt.2017.08.001.

Dong, R., Schaal, M., O'Mahony, M. P., & Smyth, B. (2013). Topic extraction from online reviews for classification and recommendation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* IJCAI '13 (p. 1310–1316). AAAI Press.

Eslami, S. P., Ghasemaghaei, M., & Hassanein, K. (2018). Which online reviews do consumers find most helpful? A multi-method investigation. *Decision Support Systems*, *113*, 32 – 42. URL: http://www.sciencedirect.com/science/article/pii/S016792361830109X. doi:10.1016/j.dss.2018.06.012.

Fan, M., Feng, C., Guo, L., Sun, M., & Li, P. (2019). Product-aware helpfulness prediction of online reviews. In *The World Wide Web Conference* WWW '19 (p. 2715–2721). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3308558.3313523. doi:10.1145/3308558.3313523.

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics. *Tourism Management*, *52*, 498 – 506. URL: http://www.sciencedirect.com/science/article/pii/S0261517715001715. doi:10.1016/j.tourman.2015.07.018.

Filieri, R., McLeay, F., Tsui, B., & Lin, Z. (2018). Consumer perceptions of information helpfulness and determinants of purchase intention in online consumer reviews of services. *Information & Management*, *55*, 956 – 970. URL: http://www.sciencedirect.com/science/article/pii/S0378720617304160. doi:10.1016/j.im.2018.04.010.

Fink, L., Rosenfeld, L., & Ravid, G. (2018). Longer online reviews are not necessarily better. *International Journal of Information Management*,

39, 30 – 37. URL: http://www.sciencedirect.com/science/article/pii/S0268401217304176. doi:10.1016/j.ijinfomgt.2017.11.002.

Fresneda, J. E., & Gefen, D. (2019). A semantic measure of online review helpfulness and the importance of message entropy. *Decision Support Systems*, *125*, 113117. URL: http://www.sciencedirect.com/science/article/pii/S0167923619301460. doi:10.1016/j.dss.2019.113117.

Gang, R., & Hong, T. (2019). Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. *Information Processing & Management*, *56*, 1425 – 1438. URL: http://www.sciencedirect.com/science/article/pii/S0306457318300360. doi:10.1016/j.ipm.2018.04.003.

Gao, B., Hu, N., & Bose, I. (2017). Follow the herd or be myself? An analysis of consistency in behavior of reviewers and helpfulness of their reviews. *Decision Support Systems*, *95*, 1 – 11. URL: http://www.sciencedirect.com/science/article/pii/S0167923616301877. doi:10.1016/j.dss.2016.11.005.

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

Ham, J., Lee, K., Kim, T., & Koo, C. (2019). Subjective perception patterns of online reviews: A comparison of utilitarian and hedonic values. *Information Processing & Management*, *56*, 1439 – 1456. URL: http://www.sciencedirect.com/science/article/pii/S0306457318300426. doi:10.1016/j.ipm.2019.03.011.

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* CSCW '00 (p. 241–250). New York,

NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/358916.358995. doi:10.1145/358916.358995.

Hernández-Rubio, M., Cantador, I., & Bellogín, A. (2019). A comparative analysis of recommender systems based on item aspect opinions extracted from user reviews. *User Modeling and User-Adapted Interaction*, *29*, 381–441. URL: https://doi.org/10.1007/s11257-018-9214-9. doi:10.1007/s11257-018-9214-9.

Hong, H., Xu, D., Wang, G., & Fan, W. (2017). Understanding the determinants of online review helpfulness. *Decision Support Systems*, *102*, 1–11. URL: https://doi.org/10.1016/j.dss.2017.06.007. doi:10.1016/j.dss.2017.06.007.

Hu, Y.-H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management*, *36*, 929 – 944. URL: http://www.sciencedirect.com/science/article/pii/S0268401215301845. doi:10.1016/j.ijinfomgt.2016.06.003.

Hu, Y.-H., Chen, K., & Lee, P.-J. (2017). The effect of user-controllable filters on the prediction of online hotel reviews. *Information & Management*, *54*, 728 – 744. URL: http://www.sciencedirect.com/science/article/pii/S0378720616304359. doi:10.1016/j.im.2016.12.009.

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, *48*, 17–27. URL: http://dx.doi.org/10.1016/j.chb.2015.01.010. doi:10.1016/j.chb.2015.01.010.

Hutto, C., & Eric, G. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (pp. 216–225). New York, NY, USA: AAAI. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

Jannach, D., Resnick, P., Tuzhilin, A., & Zanker, M. (2016). Recommender systems — beyond matrix completion. *Communication of ACM*, *59*, 94–102. URL: http://doi.acm.org/10.1145/2891406. doi:10.1145/2891406.

Jiang, M., & Diesner, J. (2016). Says who...? Identification of expert versus layman critics' reviews of documentary films. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2122–2132). Osaka, Japan: The COLING 2016 Organizing Committee. URL: https://www.aclweb.org/anthology/C16-1200.

Kim, S., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* EMNLP '06 (pp. 423–430). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=1610075.1610135.

Koren, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '08 (pp. 426–434). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1401890.1401944. doi:10.1145/1401890.1401944.

Koren, Y., & Bell, R. (2011). Advances in collaborative filtering. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 145–186). Boston, MA: Springer US. URL: https://doi.org/10.1007/978-0-387-85820-3_5. doi:10.1007/978-0-387-85820-3_5.

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2019). Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* IUI '19 (p. 379–390). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3301275.3302306. doi:10.1145/3301275.3302306.

Krestel, R., & Dokoohaki, N. (2015). Diversifying customer review rankings. *Neural Networks*, *66*, 36 – 45. URL: http://www.sciencedirect.com/

science/article/pii/S0893608015000428. doi:10.1016/j.neunet.2015.02.008.

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, *42*, 3751 – 3759. URL: http://www.sciencedirect.com/science/article/pii/S0957417414008239. doi:10.1016/j.eswa.2014.12.044.

Li, S.-T., Pham, T.-T., & Chuang, H.-C. (2019). Do reviewers' words affect predicting their helpfulness ratings? Locating helpful reviewers by linguistics styles. *Information & Management*, *56*, 28 – 38. URL: http://www.sciencedirect.com/science/article/pii/S0378720618302428. doi:10.1016/j.im.2018.06.002.

Liu, A. X., Xie, Y., & Zhang, J. (2019). It's not just what you say, but how you say it: the effect of language style matching on perceived quality of consumer reviews. *Journal of Interactive Marketing*, *46*, 70 – 86. URL: http://www.sciencedirect.com/science/article/pii/S1094996818300689. doi:10.1016/j.intmar.2018.11.001.

Loria, S. (2020). TextBlob: Simplified text processing. Https://textblob.readthedocs.io/en/dev/index.html.

Malik, M., & Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, *73*, 290 – 302. URL: http://www.sciencedirect.com/science/article/pii/S0747563217302121. doi:10.1016/j.chb.2017.03.053.

Malik, M., & Hussain, A. (2018). An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing & Management*, *54*, 88 – 104. URL: http://www.sciencedirect.com/science/article/pii/S0306457317304892. doi:10.1016/j.ipm.2017.09.004.

Margaris, D., Vassilakis, C., & Spiliotopoulos, D. (2020). What makes a review a reliable rating in recommender systems? *Information Processing & Management*, *57*, 102304. URL: http://www.sciencedirect.com/science/article/pii/S0306457320307998. doi:10.1016/j.ipm.2020.102304.

Martin, L., & Pu, P. (2014). Prediction of helpful reviews using emotions extraction. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* AAAI'14 (pp. 1551–1557). AAAI Press. URL: http://dl.acm.org/citation.cfm?id=2892753.2892768.

Mauro, N., Ardissono, L., Capecchi, S., & Galioto, R. (2020a). Service-aware interactive presentation of items for decision-making. *Applied Sciences, Special Issue Implicit and Explicit Human-Computer Interaction*, *10*, 5599. URL: https://www.mdpi.com/2076-3417/10/16/5599. doi:10.3390/app10165599.

Mauro, N., Ardissono, L., & Hu, Z. F. (2019). Multi-faceted trust-based Collaborative Filtering. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* UMAP '19 (pp. 216–224). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/3320435.3320441. doi:10.1145/3320435.3320441.

Mauro, N., Ardissono, L., & Lucenteforte, M. (2020b). Faceted search of heterogeneous geographic information for dynamic map projection. *Information Processing & Management*, *57*, 102257. URL: https://doi.org/10.1016/j.ipm.2020.102257. doi:10.1016/j.ipm.2020.102257.

Mohammadiani, R. P., Mohammadi, S., & Malik, Z. (2017). Understanding the relationship strengths in users' activities, review helpfulness and influence. *Computers in Human Behavior*, *75*, 117 – 129. URL: http://www.sciencedirect.com/science/article/pii/S074756321730225X. doi:10.1016/j.chb.2017.03.065.

Mudambi, S., & Schuff, D. (2010). What makes a helpful online review? A

study of customer reviews on Amazon.com. *MIS Quarterly*, *34*, 185–200. URL: https://doi.org/10.2307/20721420. doi:10.2307/20721420.

Nakayama, M., & Wan, Y. (2019). The cultural impact on social commerce: a sentiment analysis on Yelp ethnic restaurant reviews. *Information & Management*, *56*, 271 – 279. URL: http://www.sciencedirect.com/science/article/pii/S0378720617306225. doi:10.1016/j.im.2018.09.004.

Nelson, P. (1974). Advertising as information. *Journal of Political Economy*, *82*, 729–754. URL: https://doi.org/10.1086/260231. doi:10.1086/260231.

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: a text regression model. *Decision Support Systems*, *61*, 47 – 58. URL: http://www.sciencedirect.com/science/article/pii/S0167923614000128. doi:10.1016/j.dss.2014.01.011.

Ocampo Diaz, G., & Ng, V. (2018). Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 698–708). Melbourne, Australia: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/P18-1065. doi:10.18653/v1/P18-1065.

O'Mahony, M. P., & Smyth, B. (2010). A classification-based review recommender. *Knowledge-Based Systems*, *23*, 323–329. URL: http://dx.doi.org/10.1016/j.knosys.2009.11.004. doi:10.1016/j.knosys.2009.11.004.

O'Mahony, M. P., & Smyth, B. (2018). From opinions to recommendations. In P. Brusilovsky, & D. He (Eds.), *Social Information Access: Systems and Technologies* (pp. 480–509). Cham: Springer International Publishing. URL: https://doi.org/10.1007/978-3-319-90092-6_13. doi:10.1007/978-3-319-90092-6_13.

Pan, Y., & Zhang, J. Q. (2011). Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of Retailing*, *87*,

598 – 612. URL: http://www.sciencedirect.com/science/article/pii/S0022435911000406. doi:10.1016/j.jretai.2011.05.002.

Paul, D., Sarkar, S., Chelliah, M., Kalyan, C., & Sinai Nadkarni, P. P. (2017). Recommendation of high quality representative reviews in e-commerce. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* RecSys '17 (pp. 311–315). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3109859.3109901. doi:10.1145/3109859.3109901.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

Prado, T. R., & Moro, M. M. (2017). Review recommendation for points of interest's owners. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* HT '17 (pp. 295–304). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3078714.3078744. doi:10.1145/3078714.3078744.

Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, *20*, 542 – 556. URL: http://www.sciencedirect.com/science/article/pii/S0950705107000445. doi:10.1016/j.knosys.2007.04.004.

Qazi, A., Syed, K. B. S., Raj, R. G., Cambria, E., Tahir, M., & Alghazzawi, D. (2016). A concept-level approach to the analysis of online review helpfulness. *Computers in Human Behavior*, *58*, 75 – 81. URL: http://www.sciencedirect.com/science/article/pii/S0747563215302995. doi:10.1016/j.chb.2015.12.028.

Qi, J., Zhang, Z., Jeon, S., & Zhou, Y. (2016). Mining customer requirements from online reviews: A product improvement perspective. *Informa-*

tion & Management, *53*, 951 – 963. URL: http://www.sciencedirect.com/science/article/pii/S0378720616300581. doi:10.1016/j.im.2016.06.002.

Raghavan, S., Gunasekar, S., & Ghosh, J. (2012). Review quality aware collaborative filtering. In *Proceedings of the Sixth ACM Conference on Recommender Systems* RecSys '12 (pp. 123–130). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2365952.2365978. doi:10.1145/2365952.2365978.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 1–35). Boston, MA: Springer US. URL: https://doi.org/10.1007/978-0-387-85820-3_1. doi:10.1007/978-0-387-85820-3_1.

Robertson, S. E. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*, 503–520. URL: https://doi.org/10.1108/00220410410560582. doi:10.1108/00220410410560582.

Salehan, M., & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, *81*, 30 – 40. URL: http://www.sciencedirect.com/science/article/pii/S0167923615002006. doi:10.1016/j.dss.2015.10.006.

Shen, H. R., Rong-Ping andZhang, Yu, H., & Min, F. (2019). Sentiment based matrix factorization with reliability for recommendation. *Expert Systems with Applications*, . URL: http://www.sciencedirect.com/science/article/pii/S0957417419303951. doi:10.1016/j.eswa.2019.06.001.

Shin, D. D. (2020a). The effects of security and traceability of blockchain on digital affordance. *Online information review*, *44*, 913–932. URL: https://www.emerald.com/insight/content/doi/10.1108/

OIR-01-2019-0013/full/html?skipTracking=true. doi:10.1108/OIR-01-2019-0013.

Shin, D. D. (2020b). How do users interact with algorithm recommender systems? the interaction of users, algorithms, and performance. *Computers in Human Behavior*, *109*, 106344. URL: http://www.sciencedirect.com/science/article/pii/S0747563220300984. doi:10.1016/j.chb.2020.106344.

Shin, D. D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management*, *52*, 102061. URL: http://www.sciencedirect.com/science/article/pii/S0268401219314161. doi:10.1016/j.ijinfomgt.2019.102061.

Siering, M., Muntermann, J., & Rajagopalan, B. (2018). Explaining and predicting online review helpfulness: the role of content and reviewer-related signals. *Decision Support Systems*, *108*, 1 – 12. URL: http://www.sciencedirect.com/science/article/pii/S0167923618300149. doi:10.1016/j.dss.2018.01.004.

Sun, X., Han, M., & Feng, J. (2019). Helpfulness of online reviews: examining review informativeness and classification thresholds by search products and experience products. *Decision Support Systems*, *124*, 113099. URL: http://www.sciencedirect.com/science/article/pii/S0167923619301289. doi:10.1016/j.dss.2019.113099.

Tang, J., Gao, H., Hu, X., & Liu, H. (2013). Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems* RecSys '13 (p. 1–8). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/2507157.2507183. doi:10.1145/2507157.2507183.

Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: design and evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.),

Recommender Systems Handbook (pp. 353–382). Boston, MA: Springer US. URL: https://doi.org/10.1007/978-1-4899-7637-6_10. doi:10.1007/978-1-4899-7637-6_10.

TripAdvisor (2017). Tripadvisor. https://www.tripadvisor.it/.

Wang, J.-N., Du, J., & Chiu, Y.-L. (2020). Can online user reviews be more helpful? evaluating and improving ranking approaches. *Information & Management*, (p. 103281). URL: http://www.sciencedirect.com/science/article/pii/S0378720618310516. doi:10.1016/j.im.2020.103281.

Wu, J. (2017). Review popularity and review helpfulness: a model for user review effectiveness. *Decision Support Systems*, *97*, 92 – 103. URL: http://www.sciencedirect.com/science/article/pii/S016792361730057X. doi:10.1016/j.dss.2017.03.008.

Xiong, W., & Litman, D. (2014). Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1985–1995). Dublin, Ireland: Dublin City University and Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/C14-1187.

Xu, X., & Lu, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International Journal of Hospitality Management*, *55*, 57 – 69. URL: http://www.sciencedirect.com/science/article/pii/S0278431916300202. doi:https://doi.org/10.1016/j.ijhm.2016.03.003.

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: understanding consumer perceptions and influential factors. *International Journal of Information Management*, *37*, 673 – 683. URL: http://www.sciencedirect.com/science/article/pii/S0268401217301378. doi:10.1016/j.ijinfomgt.2017.06.004.

Yang, Y., Chen, C., & Bao, F. S. (2016). Aspect-based helpfulness prediction for online product reviews. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 836–843). IEEE. URL: https://ieeexplore.ieee.org/document/7814690. doi:10.1109/ICTAI.2016.0130.

Yang, Y., Yan, Y., Qiu, M., & Bao, F. (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)* (pp. 38–44). Beijing, China: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/P15-2007. doi:10.3115/v1/P15-2007.

Yelp (2019a). Yelp. https://www.yelp.com.

Yelp (2019b). Yelp dataset challenge. https://www.yelp.com/dataset_challenge.

Zhou, Y., Yang, S., Li, Y., chen, Y., Yao, J., & Qazi, A. (2020). Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining. *Information Processing & Management*, *57*, 102179. URL: http://www.sciencedirect.com/science/article/pii/S0306457319306788. doi:10.1016/j.ipm.2019.102179.