Modeling and Prediction of Online Product Review Helpfulness: A Survey

Gerardo Ocampo Diaz and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{godiaz,vince}@hlt.utdallas.edu

Abstract

As the popularity of free-form user-generated reviews in e-commerce and review websites continues to increase, there is a growing need for automatic mechanisms that sift through the vast number of reviews and identify quality content. Online review helpfulness modeling and prediction is a task which studies the factors that determine review helpfulness and attempts to accurately predict it. This survey paper provides an overview of the most relevant work on product review helpfulness prediction and understanding in the past decade, discusses gained insights, and provides guidelines for future research.

1 Introduction

Research on the computational modeling and prediction of online review helpfulness has generally proceeded in two directions. One concerns the automatic prediction of the helpfulness of a review, where helpfulness is typically defined as the fraction of "helpful" votes it receives. Review helpfulness research in the NLP and text mining communities has largely focused on identifying textual content features of a review that are useful for automatic helpfulness prediction. The other direction concerns understanding the nature of helpfulness, where researchers seek to understand the process of human evaluation of review helpfulness and the factors that influence it.

The increasing popularity of modeling and prediction of review helpfulness since its inception more than a decade ago can be attributed to its practical significance. Nowadays, customers regularly rely on different kinds of user reviews (e.g., hotels, restaurants, products, movies) to decide what to spend their money on. Given the large number of reviews available in web platforms, a review helpfulness prediction system could substantially save people's time by allowing them to focus on the most helpful reviews. Hence, a successful review helpfulness prediction system could be as useful as a product recommender system.

Unfortunately, unlike in many key areas of research in NLP, it is by no means easy to determine the state of the art in automatic helpfulness prediction. Empirical comparisons are complicated for at least two reasons. First, historically, systems have been trained on different datasets, not all of which are publicly available. Second, researchers have not built on the successes of each other, evaluating their ideas against baselines that are not necessarily the state of the art. Worse still, new features are not always properly evaluated. This somewhat disorganized situation can be attributed in part to the lack of a common forum for researchers to discuss a long-term vision and a roadmap for research in this area.

Our goal in this survey is to present an overview of the current state of research on computational modeling and prediction of product review helpfulness. Our focus on product reviews is motivated by the fact that they are the most widely studied type of review. Despite this focus, it is by no means the case that our work is only applicable to product reviews. While online platforms differ in objectives and review domains (e.g., Amazon is an online product store, Yelp is a business review website, and TripAdvisor is a booking website for a variety of travel activities), the principles that govern the helpfulness voting process are robust across platforms and domains. This means that most, if not all, of our findings are transferable to other kinds of online reviews. We believe that this survey will be useful to researchers and developers interested in a better understanding of the mechanisms behind review helpfulness.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 698–708
Melbourne, Australia, July 15 - 20, 2018.

2 Datasets

The main source of product reviews used in past research is Amazon.com, but interesting work has been done on data from Ciao.com (a now defunct product review website). The main difference between these two sources is the metadata associated with them: Amazon.com offers anonymous voting information, whereas Ciao attaches userIDs to helpfulness votes. Ciao also uses helpfulness votes in the range of 0 to 5, whereas Amazon votes are binary. Furthermore, Ciao offers information on a social trust network, where users choose to connect to reviewers if they find their reviews consistently helpful, unlike Amazon.com, which does not offer any such social trust network. These differences have allowed researchers to make observations on Ciao.com data that cannot be made on Amazon.com.

Datasets are collected from the aforementioned sources through web scraping or APIs. When it comes to Amazon datasets, researchers can choose one of two pre-collected datasets: the Multi-Domain Sentiment Dataset[1](Blitzer et al., 2007) (MDSD) and the Amazon Review Dataset[2] (McAuley et al., 2015; He and McAuley, 2016) (ARD). These datasets have a similar number of product categories (25 and 24, respectively). However, the latest version of MDSD contains 1,422,530 reviews, while ARD contains 142.8 million reviews. Furthermore, ARD offers a variety of metadata that is not present in MDSD (e.g., product salesrank). To the best of our knowledge, there is only one pre-collected Ciao dataset[3] (302,232 reviews, 43,666 users, and 8,894,899 helpfulness votes), which was made available by Tang et al. (2013). Few researchers have used these pre-collected datasets, however. Instead, most have relied on collecting their own datasets directly from websites. As mentioned before, the general lack of testing on pre-collected datasets has made system comparisons difficult.

The majority of researchers simply use helpfulness scores (the fraction of users who vote a review as helpful) as found in websites as ground truth for system training and evaluation. Given that these scores are volatile when reviews have few votes, researchers frequently filter out reviews that do not have a minimum number of votes. Some researchers have argued that helpfulness scores might not be good indicators of actual helpfulness, and have resorted to rating or ranking reviews themselves (Liu et al., 2007; Tsur and Rappoport, 2009; Yang et al., 2015), but these approaches are not the norm.

Votes : [97, 102]
Text : I'm a much bigger fan of the Targus folding keyboard. For starters it folds into the size of a handspring. Second of all the Landware version's keys are incredibly small. The one feature benefit of landware is that it's a rigid design so it can be used on your lap - while the Targus version is very flexible and needs to be placed on a flat surface to type.

Figure 1: Example Review

Researchers have observed interesting patterns in review datasets. For instance, positive reviews are more likely to have high helpfulness scores (O'Mahony et al., 2010; Huang et al., 2015), top ranking reviews hold a disproportionate amount of votes when compared to lower-ranked reviews (Liu et al., 2007), and more recent reviews tend to get fewer votes than older reviews (Liu et al., 2007). Although some of these effects may be the consequence of website voting mechanisms (e.g., Amazon shows reviews based on their helpfulness), they should be taken in consideration when selecting and pre-processing datasets.

Perhaps the most important observation is that helpfulness scores may not be strongly correlated to review quality (Liu et al., 2007; Danescu-Niculescu-Mizil et al., 2009; Tsur and Rappoport, 2009; Ghose and Ipeirotis, 2011; Yang et al., 2015). In at least one study, independent annotators agreed more frequently (85%) with an alternate helpfulness ranking than with one based on helpfulness scores (Tsur and Rappoport, 2009). The example review in Figure 1 shows discrepancies between quality and score. While this review is relatively short and contains only a couple of judgments on its product, 97 out of 102 people voted it as helpful (0.95 score). The quality of this review does not seem to match its near-perfect score. As we will see in Section 4, these discrepancies could be explained as the consequence of several *moderating factors*, which have a direct influence on the helpfulness voting process but are largely ignored in current helpfulness prediction systems.

[1] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/
[2] http://jmcauley.ucsd.edu/data/amazon/
[3] https://www.cse.msu.edu/~tangjili/trust.html

3 Helpfulness Prediction

Helpfulness prediction tasks include score *regression* (predicting the helpfulness score $h \in [0, 1]$ of a review), binary review *classification* (classifying a review as helpful or not), and review *ranking* (ordering a set of reviews by their helpfulness). In this section, we present the evaluation measures and approaches explored in past work.

3.1 Performance Measures

Regarding performance measures, classification tasks have used Precision, Recall, and F-measure. Regression tasks have mostly used mean squared error (MSE), which measures the average of the sum of the squared error, and root mean squared error (RMSE), which is defined as the square root of MSE. Ranking systems have used Normalized Discounted Cumulative Gain (NDCG), which is popularly used to measure the relevance of search results in information retrieval (here, helpfulness is used as a measure of relevance), and NDCG@k, a special version of NDCG that only takes into account the top k items in a ranking (this is used because users only read a limited number of reviews). Researchers have also used Pearson and Spearman correlations to measure model fit and ranking performance.

3.2 Approaches

Next, we provide a high-level overview of the approaches that have been employed to predict the helpfulness of online product reviews.

Regression has primarily been attempted through support vector regression (Kim et al., 2006; Zhang and Varadarajan, 2006; Yang et al., 2015). However, probabilistic matrix factorization (Tang et al., 2013), linear regression (Lu et al., 2010), and extended tensor factorization models (Moghaddam et al., 2012) have successfully been used to integrate sophisticated constraints into the learning process and have achieved improvements over regular regression models. Multi-layer neural networks have also been used towards this purpose (Lee and Choeh, 2014). In particular, there seems to be progress toward more sophisticated models. For instance, Mukherjee et al. (2017) used a HMM-LDA based model to jointly infer reviewer expertise, predict aspects, and review helpfulness, which showed significant improvement over simpler models. *Classification* approaches have mostly been based on SVMs (Kim et al., 2006; Hong et al., 2012; Zeng et al., 2014; Krishnamoorthy, 2015), but thresholded linear regression models (Ghose and Ipeirotis, 2011), Naive Bayes, Random Forests, J48 and JRip have also been used (O'Mahony et al., 2010; Ghose and Ipeirotis, 2011; Krishnamoorthy, 2015). Recent work has also approached this task with neural networks (Malik and Hussain, 2017; Chen et al., 2018). Regarding *ranking*, some researchers have used ranking-specific methods such as SVM ranking (Tsur and Rappoport, 2009; Hong et al., 2012), but others have attempted to recover rankings from classification (O'Mahony and Smyth, 2009, 2010) or regression (Mukherjee et al., 2017) outputs.

Table 1 provides an overview of some of the most relevant features used in helpfulness prediction systems, explains the intuition behind them and, whenever possible, their correlation to helpfulness and impact on performance. Here, we differentiate primarily between *content* and *context* features. *Content features* focus on information directly derived from the review, such as review text and star rating, whereas *context features* focus on information from outside the review, such as reviewer/user information.

Content features include *Review Length Features*, which are based on the intuition that longer reviews have more information and are thus more helpful; *Readability Features*, which are based on the conjecture that if a review is easier to read, it will be found helpful by more users; *Word-Based Features*, which are based on the idea of identifying key words whose presence indicates the importance of the information found in a review; *Word-Category Features*, which identify the presence of words belonging to specific word lists; and *Content Divergence Features*, which measure how different the contents of the review are from specific reference texts. Context features include *Reviewer Features*, which collect meaningful reviewer historical information to predict future helpfulness scores; and *User-Reviewer Idiosyncrasy Features*, which attempt to capture the similarity between users and reviewers. We also include a couple of *Miscellaneous Features*, which are based on metadata and sentiment analysis; these features are better understood in the context of the moderating factors presented in Section 4.

Researchers have managed to mostly agree on some observations regarding which features are

Feature	Description	Comments
Content Features		
Review Length Features: Measure review length using different metrics.		
Average Sentence Length	-	Used in Liu et al. (2007), Lu et al. (2010), and Yang et al. (2015) without studying its individual predictive power.
No. of Sentences	-	Used in Liu et al. (2007), Lu et al. (2010), Yang et al. (2015)
Number of Words	-	Positive correlation (Mudambi and Schuff, 2010); shown to subdue sentence features (Kim et al., 2006).
Readability Features: Measure how easy a review is to read.		
Readability	Measures how easy a text is to read	Ghose and Ipeirotis (2011) and Korfiatis et al. (2012) found a positive correlation.
Spelling Errors	-	Ghose and Ipeirotis (2011) found a negative correlation.
Paragraph Metrics	Avg. paragraph length, no. of paragraphs	Kim et al. (2006) found an insignificant difference when included in a binary classifier.
Word-Based Features: Indicate the presence of meaningful key words.		
Unigram TF-IDF	Degree of word importance in relation to all reviews for a product	Kim et al. (2006) observed a positive correlation and performance improvement when combined with review length.
Dominant Terms	Presence of particularly important terms for a specific book	Tsur and Rappoport (2009) based entire system on this metric. Tailored for book reviews: similar to UGR TF-IDF.
Word-Category Features: Indicate the presence of words of lists of semantically related words in review.		
Product features	Attempt to identify the presence of important topics	Liu et al. (2007) showed 2.89-3.22% improvement. Hong et al. (2012) presented a system which improves $\sim$ 8% accuracy over Kim et al. (2006) and Liu et al. (2007) but the individual predictive power of the feature was not analyzed. Kim et al. (2006) found it inferior to UGR TF-IDF.
Subjective Tokens	Words taken from lists of subjective adjectives and nouns	Zhang and Varadarajan (2006) found it "barely" correlated with helpfulness. No significant performance improvement.
Sentiment Words	Attempt to capture the presence of opinions, analyses, emotions etc.	Kim et al. (2006) found these features inferior to UGR TF-IDF; Yang et al. (2015) found the opposite and significant improvement over simple text features regression.
Syntactic tokens	A variety of tokens including nouns, adjectives, adverbs, wh- determiners etc.	Kim et al. (2006) found no performance gains; Hong et al. (2012) built a system with volition auxiliaries and sentence tense which showed $\sim$ 8% accuracy improvement over Kim et al. (2006) and Liu et al. (2007), but the individual predictive power of these features was not studied.
Content Divergence Features: Measure the difference between reviews and some reference text.		
Review-product descr. divergence	Helpful reviews should echo the contents of product description	Zhang and Varadarajan (2006) found no significant improvement in model correlation.
Sentiment divergence	The mainstream opinion polarity for a product and its strength are compared to those of the review	Hong et al. (2012) presented a system which improved $\sim$ 8% accuracy over Kim et al. (2006) and Liu et al. (2007) but the individual predictive power of the feature was not analyzed.
KL average review divergence	Divergence between the unigram language model of the review and aggregated product reviews	Lu et al. (2010) introduced it in their baseline model along with a variety of features; the individual predictive power of the feature was not studied.
Miscellaneous Features		
Star rating	The review-assigned product star rating	Positively correlated to helpfulness (Huang et al., 2015). Influence explained by Danescu-Niculescu-Mizil et al. (2009) and Mudambi and Schuff (2010) (see Sections 4.4, 4.2).
Subjectivity	The probability of a review and its sentences being subjective	Based on the conjecture that readers prefer subjective or objective info. based on product type. Empirical evidence found in Ghose and Ipeirotis (2011) (see Section 4.5).
Context Features		
Reviewer Features: Capture reviewer statistics.		
# Past Reviews	Previous reviews written by reviewer	No influence found by Huang et al. (2015).
# Helpful Votes	Previous votes received by reviewer	No influence found by Huang et al. (2015).
Avg. Helpfulness	Reviewer avg. past helpfulness	Positive correlation found by Huang et al. (2015). Mixed effects found by Ghose and Ipeirotis (2011).
User-Reviewer Idiosyncrasy: Capture the similarity between users and reviewers.		
Connection Strength	User-Reviewer connection strength in a social network using the metric introduced in Tang et al. (2012)	Relative performance increase of 1.15-28.38% (Lu et al., 2010; Tang et al., 2013) (see Section 4.3)
User-Reviewer Product Rating Similarity	User-Reviewer product rating history similarity	Relative performance increase of 28.38% (Tang et al., 2013) (see Section 4.3)

Table 1: Summary of Observed Features on Helpfulness

useful for helpfulness prediction[4]. *Review length* has been shown multiple times to be strongly (positively) correlated to helpfulness (Kim et al., 2006; Liu et al., 2007; Otterbacher, 2009; Mudambi and Schuff, 2010; Cao et al., 2011; Pan and Zhang, 2011; Yang et al., 2015; Bjering et al., 2015; Huang et al., 2015; Salehan and Kim, 2016) with only few researchers disagreeing on the existence of the correlation (Zhang and Varadarajan, 2006; Korfiatis et al., 2012). There is general agreement that a review's *star rating* can also be useful for helpfulness prediction. Some researchers use the extremity of the rating (positive, negative, neutral) as a feature (positive and negative reviews are seen as more useful than neutral reviews) (Ghose and Ipeirotis, 2011), while others use star ratings directly (Kim et al., 2006; Mudambi and Schuff, 2010; Pan and Zhang, 2011; Zeng et al., 2014; Huang et al., 2015; Bjering et al., 2015). Some researchers argue that star rating is useful because of the presence of *positivity bias* (i.e., reviews with positive star ratings are seen as more helpful), while few researchers disagree on the existence of a connection between star ratings and helpfulness (Otterbacher, 2009). *Review readability metrics*, which measure how "easy" it is to read a review, have been found to have a positive correlation to helpfulness (Ghose and Ipeirotis, 2011; Korfiatis et al., 2012), but have not been as thoroughly tested as other features. A recurrent idea is that of capturing *review content relevance*: unigram TF-IDF statistics (the relative importance of the words in a review when compared to other reviews of the same product) (Kim et al., 2006), dominant terms (computed using a custom metric similar to TF-IDF, but tailored for book reviews) (Tsur and Rappoport, 2009), and latent review topics (the themes present in the review) (McAuley and Leskovec, 2013; Mukherjee et al., 2017) stand out particularly.

3.3 The State of Helpfulness Prediction

The classical approach to helpfulness prediction has consisted of finding new hand-crafted features that can improve system performance. Although many interesting features continue to be found (e.g., emotion (Martin and Pu, 2014), aspect (Yang et al., 2016), and argument (Liu et al., 2017) based features), advances have been hindered by the lack of standard datasets, which are needed for performance comparisons, and feature ablation studies, which are needed to properly evaluate the contribution of newly proposed features.

Even so, as in many other areas of NLP, recent systems based on neural network architectures have shown performance increases both when using hand-crafted features (Lee and Choeh, 2014; Malik and Hussain, 2017) and when performing raw-text predictions (Chen et al., 2018). Moreover, recent systems have been shown to be able to tackle domain knowledge transfer considerably well (Chen et al., 2018). Although these systems were not compared against a robust hand-crafted feature baseline, the fact that authors are beginning to use pre-collected datasets (ARD) enables fairer comparisons. Intuitively, we expect models based on neural network architectures to be better at capturing latent semantics, as well as some of the feature interactions we will present in Section 4. In parallel, systems that have incorporated user and reviewer features, particularly those that learn from individual user votes (Tang et al., 2013), have shown large performance increases over extensive hand-crafted-only feature baselines (Lu et al., 2010; Tang et al., 2013), and more sophisticated models focused on review semantics (Mukherjee et al., 2017) have also outperformed hand-crafted-only feature baselines significantly.

4 The Helpfulness Voting Process: Entities and Moderating Factors

So far we have presented an overview of the features used in helpfulness prediction systems. With a few exceptions (Mudambi and Schuff, 2010; Ghose and Ipeirotis, 2011; Tang et al., 2013), past work on helpfulness prediction has focused exclusively on *non-moderating factors* (i.e., observable features which can contribute towards helpfulness scores, but cannot alter or influence the voting process itself). Even so, researchers have gained key insights on certain *moderating factors* (i.e., mechanisms and properties that can influence the voting process outcome). These findings are relevant not only because they can be used to enhance helpfulness prediction, but because, when put together, they constitute arguments in favor of reconsidering the helpfulness prediction task and its focus. In this section, we will present a variety of moderating factors.

[4]We do not discuss features that are not helpful since, in general, they are not as thoroughly tested as those mentioned here.

4.1 The Voting Process and its Entities

To start our discussion on moderating factors, let us provide a brief, intuitive definition of the steps involved in the helpfulness voting process and outline the entities involved in it[5]:

1. A reviewer, a, writes a review r on product p
2. A user, u, reads the review by reviewer a on product p and internally assigns it a score s using some criterion c.
3. If the score s is over some threshold t, the user votes the review as "helpful". Otherwise, the user votes it as "not helpful".

Intuitively, one can expect these four entities — reviewers, users, reviews, and products — to play a role in determining the outcome of the voting process. Moreover, it is reasonable to expect both the nature of these entities and the interactions between them to be sometimes expressed through hidden features/variables. For instance, one cannot directly observe a user's opinion of a product unless he/she writes a review, and one cannot directly observe a particular user's information needs or a product's nature, which would indicate what kind of review is most helpful for it. In the next subsections, we will discuss different moderating factors that have been discovered for each of these entities, the observable features that have been used to approximate them, and their effects on the voting process.

4.2 User-Product Predispositions

Danescu-Niculescu-Mizil et al. (2009) showed that the difference between user and reviewer opinions can influence helpfulness votes. Since user opinions are hidden, based on the assumption that star ratings are good indicators of opinion, Danescu et al. studied the interplay between review star rating deviation from the mean (the divergence between the reviewer's opinion and the average opinion of the product) and star rating variance (the level of opinion consensus for a product) for 1 million Amazon US book reviews, making the following observations:

1. When star rating variance is very low, the most helpful reviews are those with the average star rating.
2. With moderate variance, the most helpful reviews are those with a slightly-above-average star rating.
3. As variance becomes large, reviews with star ratings both above and below the average are more helpful (positive reviews still deemed somewhat more helpful).

These observations held when controlling for review text, and constitute one of the most straightforward pieces of evidence against text-only review helpfulness understanding and prediction. Although these observations show only aggregated user behavior, they have a theoretical backing by past research (Wilson and Peterson, 1989), and hint that a deeper understanding of user opinions can lead to better prediction systems.

4.3 User-Reviewer Idiosyncrasy

Tang et al. (2013) found that, by observing users' actions, user-reviewer idiosyncrasy similarity could be measured and used to enhance helpfulness prediction. They showed that the existence and strength of connections between reviewers and users in a social network, along with product rating history similarity, moderated the general user opinion of a particular reviewer's reviews. Specifically, they analyzed *social network connections* in Ciao's *circle of trust*, a social network where a user *connects* to a reviewer if they consistently find their reviews helpful, along with users' and reviewers' product rating histories, and made the following observations:

1. Users are likely to think of reviews from their connected reviewers as more helpful.
2. The more strongly users connect to a reviewer, the more helpful users consider the reviews from the reviewer.[6]
3. Users are likely to consider the reviews from reviewers with similar product ratings as more helpful.
4. The more similar the product ratings of users and reviewers, the more helpful users consider the reviews from the reviewer.

As Tang et al. proposed that differences in helpfulness scores are not necessarily a consequence of review quality, but of differences of opinion between users (if everyone thought the same way, all reviews would have a score of either 0 or 1), they were among the first to advocate for user-specific helpfulness prediction, which aims to predict how a specific user will vote, instead of predicting the

[5]Here we assume voting participation and do not attempt to reconcile it with polarity, but a deeper understanding of participation could lead to better interpretations of votes.

[6]Connection strength is measured with the metric introduced in Tang et al. (2012).

aggregated votes of the community. Under this approach, Tang et al. implemented their observations in a probabilistic matrix factorization framework and achieved a 28.38% relative improvement over a text-reviewer-based baseline that included an extensive set of text features present in other systems (Lu et al., 2010).

This suggests that the similarity between reviewers' idiosyncrasy as expressed in reviews and that of users can be approximated by studying user and reviewer actions. Further, the information used by Tang et al. (2013) towards this purpose is not the only kind that could prove useful. It could easily be extended to include the vast amount of user information stored by current day e-commerce websites such as Amazon. Users' age, gender, purchase history, location, browsing and purchase patterns, and review history (both writing and rating) could be used to define prior probabilities on some user x liking the review of a reviewer y.[7] As some of this information has already been used in recommender systems, it would be of interest to explore the extent to which techniques from this field (specifically those from collaborative filtering) can be applied to helpfulness prediction.

4.4 Product Nature

Product nature moderates users' information needs and the criteria of a helpful review. Online stores now have an astoundingly large catalog of products, which can be very different in price, use, target market, complexity, popularity, etc. Hence, it is reasonable to expect the information needs of users to depend at least somewhat on the product in question. Consider the task of *buying a house vs buying a TV*. We can easily see that the amount and nature of information needed to buy a TV or a house is considerably different. Further, the quality of these products stems from different sources: a TV's perceived quality depends mostly on its technical features, whereas the perceived quality of a house depends to some degree on the potential buyer. Therefore, it is perfectly sensible to expect helpful reviews for products of different "types" to be different. Below we show that the nature of a product moderates the effects of star ratings, review length, and subjectivity on helpfulness scores.

Researchers have proven the influence of product nature on the helpfulness voting process by differentiating between *search* and *experience* goods. According to Nelson (1970, 1974), the quality of search goods is derived from objective attributes (e.g., a camera), whereas the quality of experience goods is based on subjective attributes (e.g., a music CD). Mudambi and Schuff (2010) first identified that review length (word count) is positively correlated to review helpfulness, and then made the following observations:

- For experience goods, reviews with extreme star ratings (high or low) are associated with lower levels of helpfulness than reviews with moderate star ratings.
- Review depth has a greater positive effect on the helpfulness of the review for search goods than experience goods.

These observations make it clear that the nature of a product can impact the way a user will judge a review's helpfulness. However, approximating the nature of a product is not a trivial task. As stated by Mudambi and Schuff, even if these observations hold, classifying products as search or experience goods is a complicated task, since products fall at some point along a spectrum and commonly have aspects of both search and experience goods. This means that finding methods of automatically discovering product features or classifications that influence the helpfulness voting process is an important task for future research.

What other product categorizations are there that could influence helpfulness and be easily collected/computed? We propose to start by using categories already present in e-commerce websites. Intuitively, it would make sense for products under the "computers" category to be similar in their information needs. And as such, systems trained on computer reviews should learn similar parameters. As most e-commerce websites use a hierarchical product categorization system, by starting at the most specific subcategories one could potentially generalize subcategory-learned parameters into category-wide trends.

4.5 Review Nature

A review's style influences the properties that make it helpful. It is well known that when it comes to expressing opinions, the way information is presented can be almost as important as the

[7] Since a reviewer's idiosyncrasy is embodied in his/her reviews, we do not rule out the possibility that more complex text representations can also be used to approximate it. Regardless, these sources of information should still be able to complement prediction systems.

information itself. Even if two reviewers have a similar opinion on a product, the way they frame their opinion can make a big difference when it comes to how helpful their reviews are. Consider the task of deciding whether to buy a specific car. What advice could prove useful for this decision? We could consider *regular* advice that is mostly concerned with the car itself, *comparative* advice that relates various aspects of the car with its alternatives, and *suggestive* advice, which focuses on usage recommendations.

Qazi et al. (2016) used these three types of advice to classify hotel reviews from TripAdvisor.com and made the following observations:

- For comparative reviews, longer reviews are considered more helpful.
- For suggestive and regular reviews, shorter reviews are more helpful.

Similar findings on the influence of review nature were made by Huang et al. (2015): when differentiating between reviews written by regular and top Amazon reviewers, they made the following observations:

- The influence of word count on review helpfulness is bounded (after 144 words, the effect stops) for regular reviewers.
- For top reviewers, the effect is nonexistent.

Similarly to product nature, an important research question for future work is how to identify and exploit review categories for effective helpfulness prediction. We expect more sophisticated textual features to be necessary to differentiate between meaningful styles of reviews.

4.6 Review Context

Sipos et al. (2014) found evidence that helpfulness votes are the consequence of judgments of *relative quality* (i.e., how the review compares to its neighbors) and that aggregate user voting polarity is influenced by the specific review ranking that websites display at any given point in time. To prove this, they collected daily snapshots of the top 50 reviews of 595 Amazon products over a 5 month period. Four months after the data collection period ended, they collected the full review rankings for all 595 products. This final review ranking was taken to be the "true" ranking. They studied daily changes and observed that:

- A review receives more positive votes when it is under-ranked (under its final ranking).
- A review receives more positive votes when it is superior to its neighbors.
- A review receives fewer positive votes when it is over-ranked (over its final ranking).
- A review receives fewer positive votes when it is locally inferior to its neighbors.

Sipos et al. noted that these observations are consistent with the interpretation that users vote to correct "misorderings" in the ranking. This has important consequences for user-specific helpfulness prediction systems. Recall that votes may express judgments over a set of reviews. If researchers build training sets that identify user votes and contain sufficient information to replicate context at the time of voting, systems could learn more about user preferences: a vote would no longer inform solely on a user's perceived helpfulness of a review x, but on the user's perceived helpfulness of x with respect to its neighbors. This could be particularly useful in sparsity scenarios, and could lead to better helpfulness predictions.

5 Conclusions and Recommendations

Online product review helpfulness modeling and prediction is a multi-faceted task that involves using content and context information to understand and predict helpfulness scores. Researchers now have at their disposal at least three public, pre-collected product review datasets — MDSD, ARD, and Ciao — to build and test systems. Although significant advances have been made on finding hand-crafted features for helpfulness prediction, effective comparisons between proposed approaches have been hindered by the lack of standard evaluation datasets, well-defined baselines, and feature ablation studies. However, there have been exciting developments in helpfulness prediction: systems that have attempted to exploit user and reviewer information, along with those based on sophisticated models (e.g., probabilistic matrix factorization, HMM-LDA) and neural network architectures, are promising prospects for future work. Furthermore, a variety of insightful observations have been made on moderating factors. In particular, product opinions, user idiosyncrasy, product and review nature, along with review voting context have been shown to influence the way users vote. This provides suggestive evidence that researchers should adopt a holistic view of the helpfulness voting process, which may require information not present in current datasets.

We conclude our survey with several recommendations for future work on computational modeling and prediction of review helpfulness.

Task If one acknowledges the role that users play in determining whether a review is helpful or not, it seems contradictory to insist on predicting helpfulness scores, which represent the average perception of a subset of users that (1) may not be representative of the entire population and (2) may not serve users well if their perceptions do not align with the subset of users that voted (even if the subset consisted of the entire population). This is why we consider that user-specific helpfulness prediction, first presented in Moghaddam et al. (2012) and Tang et al. (2013), should be the goal of future work, as it allows systems to tailor their predictions to users' preferences and needs (much like a recommender system). Note that pursuing user-specific helpfulness prediction is not enough. A substantial amount of work must still be done to find, approximate, and implement moderating factors in helpfulness prediction systems, as well as build models that can adequately reflect the effects of these factors.

Data Given that we recommend user-specific helpfulness prediction, we propose the development of a gold standard that contains information that can facilitate the design of user-specific models (e.g., records of who voted and how, data relevant to user-profiling recommendations such as age, location, social networks, purchase and browsing history and patterns, product reviews written, and review and product rating histories). Furthermore, as users frequently vote on reviews in a different context (scores and neighboring reviews can vary over time), this dataset should include temporal information, which would allow researchers to reconstruct the context under which votes are cast. To build this dataset, we recommend that researchers work with companies such as Amazon, which may have such information.

Features and knowledge sources While we encourage the development of user-specific helpfulness prediction, we by no means imply that a model should be trained for each user. In fact, this may not be feasible if a user has cast only a small number of votes. There are multiple ways to approach this task. One is to train a user-specific model for each *cluster* of "similar" users. Taking inspirations from collaborative filtering, we could define or learn user similarity based on their purchasing/browsing/review and product rating histories (Liu et al., 2014) as well as profiling information (Krulwich, 1997), which should be available in the aforementioned dataset. Further, "similar" reviews (i.e., reviews on which users vote similarly) could be exploited (Sarwar et al., 2001; Linden et al., 2003). Once product and user/reviewer factors are incorporated into a model, it may become feasible to use past instances to predict helpfulness votes (how similar is a test instance to past situations where a user has voted "helpful"?).

Baseline systems To design a strong baseline system, first, researchers should consider all proposed features so far, including content features, context features, and features used to approach moderating factors. Second, combinations of these features should be systematically tested on the different models proposed by researchers. As we have seen that product nature influences the voting process, these tests should be conducted over different products and product categories. We recommend identifying specific experience and search products, since the effects of product nature have already been proven for them. Although ideally, these tests would be carried out on our proposed gold-standard dataset, we believe that the Ciao dataset introduced in Tang et al. (2013) and ARD (McAuley et al., 2015) can prove useful to define a baseline in the short term. Towards this purpose, the systems proposed in Tang et al. (2013), Mukherjee et al. (2017), Malik and Hussain (2017), and Chen et al. (2018) could serve as baselines after being enriched with extra features.

Other platforms, review domains and languages While we focused on Amazon product reviews written in English, the majority of the features discussed in Section 3 are platform-, domain- and language-independent, and the existence and importance of moderating factors described in Section 4 is by no means limited to product reviews. Consequently, we encourage researchers to evaluate the usefulness of these features and study these moderating factors in different domains, platforms, and languages, possibly identifying new features and moderating factors.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was supported in part by USAF Grant FA9550-15-1-0346.

References

Einar Bjering, Lars Jaakko Havro, and Oystein Moen. 2015. An empirical investigation of self-selection bias and factors influencing review helpfulness. *International Journal of Business and Management*, 10(7):16–30.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.

Qing Cao, Wenjing Duan, and Qiwei Gan. 2011. Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. *Decision Support Systems*, 50(2):511–521.

Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. 2018. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 602–607.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on Amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, pages 141–150.

Anindya Ghose and Panagiotis G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 507–517.

Yu Hong, Jun Lu, Jianmin Yao, Qiaoming Zhu, and Guodong Zhou. 2012. What reviews are satisfactory: Novel features for automatic helpfulness voting. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 495–504.

Albert H. Huang, Kuanchin Chen, David C. Yen, and Trang P. Tran. 2015. A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, 48:17–27.

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*, pages 423–430.

Nikolaos Korfiatis, Elena García-Bariocanal, and Salvador Sánchez-Alonso. 2012. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217.

Srikumar Krishnamoorthy. 2015. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.

Bruce Krulwich. 1997. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI Magazine*, 18(2):37–45.

Sangjae Lee and Joon Yeon Choeh. 2014. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046.

Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, and Xuzhen Zhu. 2014. A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems*, 56:156–166.

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*, pages 1358–1363.

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*, pages 334–342.

Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th International Conference on World Wide Web*, pages 691–700.

M.S.I. Malik and Ayyaz Hussain. 2017. Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, 73:290–302.

Lionel Martin and Pearl Pu. 2014. Prediction of helpful reviews using emotions extraction. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1551–1557.

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*, pages 165–172.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52.

Samaneh Moghaddam, Mohsen Jamali, and Martin Ester. 2012. ETF: Extended tensor factorization model for personalizing prediction of review helpfulness. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 163–172.

Susan M. Mudambi and David Schuff. 2010. What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, 34(1):185–200.

Subhabrata Mukherjee, Kashyap Popat, and Gerhard Weikum. 2017. Exploring latent semantic factors to find useful product reviews. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 480–488.

Phillip Nelson. 1970. Information and consumer behavior. *Journal of Political Economy*, 78(2):311–329.

Phillip Nelson. 1974. Advertising as information. *Journal of Political Economy*, 82(4):729–754.

Michael P. O'Mahony, Pádraig Cunningham, and Barry Smyth. 2010. An assessment of machine learning techniques for review recommendation. In *Artificial Intelligence and Cognitive Science*, pages 241–250.

Michael P. O'Mahony and Barry Smyth. 2009. Learning to recommend helpful hotel reviews. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 305–308.

Michael P. O'Mahony and Barry Smyth. 2010. A classification-based review recommender. *Knowledge-Based Systems*, 23(4):323–329.

Jahna Otterbacher. 2009. Helpfulness in online communities: A measure of message quality. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 955–964.

Yue Pan and Jason Q. Zhang. 2011. Born unequal: A study of the helpfulness of user-generated product reviews. *Journal of Retailing*, 87(4):598–612.

Aika Qazi, Karim Bux Shah Syed, Ram Gopal Raj, Erik Cambria, Muhammad Tahir, and Daniyal Alghazzawi. 2016. A concept-level approach to the analysis of online review helpfulness. *Computers in Human Behavior*, 58:75–81.

Mohammad Salehan and Dan J. Kim. 2016. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, 81:30–40.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295.

Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. 2014. Was this review helpful to you?: It depends! Context and voting patterns in online content. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 337–348.

Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 1–8.

Jiliang Tang, Huiji Gao, and Huan Liu. 2012. mTrust: Discerning multi-faceted trust in a connected world. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 93–102.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *International AAAI Conference on Web and Social Media*, pages 154–161.

William R. Wilson and Robert A. Peterson. 1989. Some limits on the potency of word-of-mouth information. *Advances in Consumer Research*, 16:23–29.

Yinfei Yang, Cen Chen, and Forrest Sheng Bao. 2016. Aspect-based helpfulness prediction for online product reviews. In *Proceedings of the 28th IEEE International Conference on Tools with Artificial Intelligence*, pages 836–843.

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)*, pages 38–44.

Yi-Ching Zeng, Tsun Ku, Shih-Hung Wu, Liang-Pu Chen, and Gwo-Dong Chen. 2014. Modeling the helpful opinion mining of online consumer reviews as a classification problem. *International Journal of Computational Linguistics & Chinese Language Processing*, 19(2):17–32.

Zhu Zhang and Balaji Varadarajan. 2006. Utility scoring of product reviews. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 51–57.