Was This Review Helpful to You? It Depends! Context and Voting Patterns in Online Content

Ruben Sipos
Dept. of Computer Science
Cornell University
Ithaca, NY
rs@cs.cornell.edu

Arpita Ghosh
Dept. of Information Science
Cornell University
Ithaca, NY
arpitaghosh@cornell.edu

Thorsten Joachims
Dept. of Computer Science
Cornell University
Ithaca, NY
tj@cs.cornell.edu

ABSTRACT

When a website hosting user-generated content asks users a straightforward question — "Was this content helpful?" with one "Yes" and one "No" button as the two possible answers — one might expect to get a straightforward answer. In this paper, we explore how users respond to this question and find that their responses are not quite straightforward after all. Using data from Amazon product reviews, we present evidence that users do not make absolute, independent voting decisions based on individual review quality alone. Rather, whether users vote at all, as well as the polarity of their vote for any given review, depends on the *context* in which they view it — reviews receive a larger overall number of votes when they are 'misranked', and the polarity of votes becomes more positive/negative when the review is ranked lower/higher than it deserves. We distill these empirical findings into a new probabilistic model of rating behavior that includes the dependence of rating decisions on context. Understanding and formally modeling voting behavior is crucial for designing learning mechanisms and algorithms for review ranking, and we conjecture that many of our findings also apply to user behavior in other online content-rating settings.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance Feedback; H.3.5 [**Online Information Services**]: Web-Based Services

General Terms

Human Factors

Keywords

User-generated Content, Online Reviews, User Feedback, Human Computation, Ratings, Ranking

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2567998.

1. INTRODUCTION

User-generated content (UGC) is now one of the primary sources of useful content on the Web. But while there is a tremendous volume of it — thanks to a lack of barrier to contribution — not all of it is equally good. This means that sorting and ranking content is essential to making UGC actually useful to a site's viewers. To this effect, most sites with user-generated content — such as reviews on Amazon, answers on online Q&A sites like StackOverflow, articles on Digg or Reddit, or comments on news articles and YouTube videos — allow viewers to rate content and use these ratings to determine the order in which to display contributions.

In an ideal world, users would respond to questions about rating — such as Amazon's "Was this review helpful to you? Yes/No" query — by judging each contribution exclusively based on its absolute merits, independent of the contribution's display context or its previous ratings. But does real user voting behavior resemble this ideal?

Addressing this question and understanding user rating behavior is important for more than one reason. First, when contributions are displayed in order of aggregate user ratings (such as the ratio of "yes" votes to total votes), whether higher quality contributions are indeed ranked higher or not depends on *what a vote actually means*: if user voting behavior is such that votes are unfavorably biased by other factors, the resulting ranking may not be the one the website seeks. Second, an accurate model of user behavior is important for designing optimal algorithms to quickly *learn* contributions' qualities from votes. In particular, the performance of learning algorithms based on user ratings will depend on correctly interpreting the information conveyed by the votes in the first place.

In this paper, we address the question of how users rate content, using data collected over 5 months on 595 products from Amazon, with daily statistics on the received votes for each review in this set of products. We find that an *absolute* rating model, where users cast votes that depend purely on judging a review's merit in isolation, is inaccurate and does not fit observed voting patterns. Instead, users appear to cast votes that reflect *relative*, rather than absolute, judgments about reviews' quality. In particular, we find that both *how* a user rates the review, as well as *whether* a user votes on it at all, varies with the review's context — the relative quality of the surrounding reviews, as well as its current ranking due to ratings from previous voters.

We note here that ratings on Amazon reviews provide a relatively "neutral" sample of user-generated content, which makes them a good environment for understand voting be-

havior. In other rating environments, such as comments on news articles or answers on Q&A sites, votes could (and are anecdotally known to) indicate not only content quality, but also "Agree/Disagree" or "I like/dislike your opinion". In contrast, reviews on Amazon — at least for the vast majority of non-controversial products[1] — are rated by users *before* the user has experienced the product (since users would typically read reviews on Amazon to help decide whether or not to purchase a product), so that thumbs-up/down ratings of reviews on Amazon are more likely to relate to the review's quality rather than to reflect agreement or differences of opinion with it.

Organization. We first describe our dataset in Section 2. Our empirical analysis of the data starts in Section 3 with a statistical analysis that demonstrates that users do not vote according to absolute independent judgments alone. We then explore how context relates to voting polarity and participation in Section 4 and Section 5 respectively. We conclude with a discussion in Section 6 and related work in Section 7.

2. DATASET

To understand how users vote on online content, we need a dataset that contains information on how votes are cast. Publicly available datasets on user-generated content are typically snapshots of a particular site (such as Amazon) at a specific point in time, containing information about the content on the site at that point, current rankings of the content (*i.e.*, the order in which contributions are displayed), and cumulative ratings (e.g., how many users found each contribution helpful during its lifetime up until the snapshot). While this allows reasoning about, for example, how the content of a contribution influences the votes it receives, it is not sufficient to address voting behavior — when and why the contributions accumulated the votes they did.

This led us to collect our own dataset, which is scraped from publicly viewable data on Amazon. For every product, Amazon displays a list of all its reviews, as well as the accumulated votes on each review — how many Amazon users rated that particular review helpful (or not), up-to that point. We wrote a python script that retrieved and parsed these web pages to obtain a sequence of snapshots that contained (for every product) the list of reviews, in what order Amazon displayed these reviews (*i.e.*, their current rankings), and the number of "yes" and "no" votes for each review.

We selected a set of 595 products[2] chosen from the top 100 products of six Amazon "Hot New Releases" lists (Books, Video Games, Music, Movies & TV, Toys & Games, Electronics) as of 2nd October 2012. These products were tracked *daily* for a period of 5 months from October 2012 to March 2013. Over this period, we periodically ran our script to collect data on the 50 most helpful reviews (or fewer, if there are less than 50 reviews) for each of these products. Our choices in data collection are driven by the following reasons. (i) First, we chose a subset of products rather than all products because tracking all products on Amazon would be unjustifiably resource-intensive. (ii) To ensure adequate data points for each review, we focused on popular products which receive more reviews than niche products[3]. (iii) Finally, to include early votes on early reviews on products in our dataset, we would like to follow products from 'the beginning', *i.e.*, from the time of their launch on Amazon or close to it. We used Amazon's "Hot New Releases" lists as a proxy for future popularity, to address the problem of choosing a subset of popular products which also allow observing the early voting dynamics: these lists contain new (just or soon-to-be released) products that Amazon expects to be popular enough to satisfy the criterion for (predicted) high reviewing and voting volume. The products we track were all selected at the same time, which was the beginning of our data collection period. We restrict ourselves to no more than 50 top reviews for each product due to politeness concerns, since there are a small number of popular products that are collecting thousands of reviews — a single pass using a reasonable delay between HTTP requests already required over an hour, even when restricted to 595 products with at most 50 reviews per product.

The interval between the script runs was one day (although the time of day when the script ran was not fixed, and changed over time). This lead to a dataset containing 150 daily snapshots, 71504 reviews and 497088 votes. The key feature of this dataset is that it allows us to study how votes were cast over time, and as a function of the context in which the review was viewed. In particular, we can see how many votes a review received on a particular day, its rank on that day, and how other reviews were ranked in relation to it.

We note some important caveats about our data, however. (i) Our data only consists of daily snapshots and not *individual* user interactions, so that users, and rankings, are aggregated at the daily level. Specifically, this means that if the ranking changed more often than once per day, we have a mismatch between our data and some actual users' experiences, which can blur our measurements. (ii) We do not have a way of tracking page views. This is relevant to our analysis of participation, *i.e.*, when users choose to vote; see Section 5. (iii) Finally, we have a few instances of incorrect parsing due to changes in Amazon's webpage structure, which needed to be manually corrected for the final dataset. Nevertheless, there remain rare instances of mismatched data as a natural consequence of the temporal nature of the data we seek (since the script could not be rerun to correct any given snapshot after the corresponding day had passed).

Filtering. We filter our dataset to focus on voting behavior on the "*average* review", eliminating all (day, review) pairs where a review received more than 10 positive or negative votes between consecutive snapshots (except for the analysis described in Section 4.1 and Section 5.1). This is to prevent such rare, very popular reviews from dominating and overshadowing the overall pattern of voting on reviews, since most reviews — even for our subset of popular products — receive at most a few votes (if any) between consecutive snapshots (*i.e.*, within a single day interval). We

[1]Some books on Amazon do appear to have highly polarized reviews.

[2]Five items in two lists did not parse correctly and were automatically excluded.

[3]Note that this means that our observations regarding voting behavior are possibly only valid for reviews on popular products. However, this is arguably the most relevant set of reviews for which to study voting behavior, since reviews on frequently-purchased products are likely the most useful category of reviews on Amazon.

note that this means that our results are likely *not indicative* of voting behavior on the extremely popular reviews which might have interestingly different voting patterns. Our focus here is on the remaining datapoints (after filtering we retain >90% of datapoints and >60% of votes) which represent the more typical review on Amazon (for our dataset the average number of votes in a day is less than 2 even if we exclude days with no votes). We note, however, that the trends observed in the plots in Section 5 and Section 4 can still be seen when this data is not filtered out, albeit with more noise.

2.1 Rating Accumulation Process

We now discuss how reviews and ratings accumulate as a function of time elapsed following a product's release. Figure 1 shows that the rate at which new reviews appear remains roughly constant for almost the first 4 months of our 5-month data collection period, and remains non-negligible, although clearly lower, even in the last month. The plot does not start at zero reviews because the lists of products on "Hot New Releases" can also contain already released products, so that some very early reviews may already be present at $t = 0$. Also, since we scrape only the first 50 reviews for each product, any new reviews that appear below the 50th rank are ignored (unless and until they place in the top 50) — so this plot is essentially a lower bound on the actual total number of reviews[4].

Similarly, new votes appear throughout our collection interval (Figure 1). Again, the rate slows down over time but remains substantially larger than zero. The kinks in this plot arise from *cross-listing* — a jump occurs when Amazon cross-lists a review (along with its votes) from one product (not in our dataset) to another similar one (that we track) on that day. Figure 1 suggests that the voting patterns we observe are not particularly influenced by the "age" of the reviews, in the sense that most of the reviewing and voting does not occur only very early in our data set[5].

2.2 Converged Rankings

Approximately 4 months after the end of our data collection period, we collected the full ranking of all reviews for all products in our dataset (*i.e.*, not restricting it to the top 50 reviews as for the daily snapshots). We refer to the rank of a review in this final snapshot as its *final true rank*, accounting for ties[6].

Comparing these final true ranks to the observed ranks towards the end of our data collection period shows that the top 30 positions of the rankings are largely stable by this time. In particular, Figure 2(b) shows that during the last 30 days of our data collection period, most reviews are already consistently close to the ranking positions that we observed 4 months later (after removing reviews that were published after our data collection period ended). We therefore conclude that the ranking process converges and that the relative ordering of reviews stabilizes.

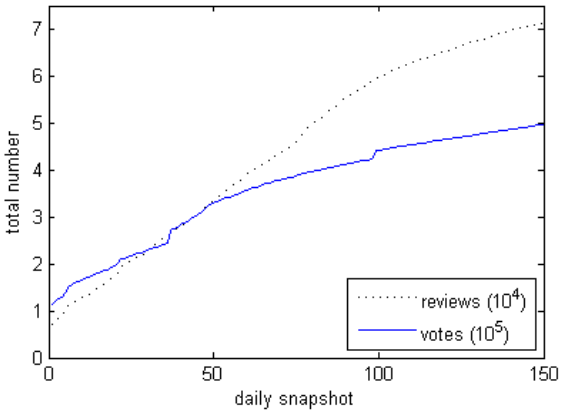


Figure 1: Total number of reviews and votes (y-axis) over time (x-axis in days).

To sanity-check that this convergence is not a pathology of the (secret) ranking algorithm that Amazon uses (e.g., Amazon may just decide to fix the ordering after a few months), we empirically investigated how Amazon computes a ranking in response to the helpfulness votes. In particular, we consider a smoothed version of the ratio of "yes" votes to total votes as the ranking criterion (specifically, ranking by "yes" votes divided by total number of votes plus one). This simple ranking criterion achieves a Kendall-tau rank correlation coefficient (with ties) of 0.84 (using only reviews with at least 10 total votes and products with at least 2 reviews). While we do not obtain a perfect correlation (Amazon may use factors beyond user ratings, exploration strategies, and possibly more complex functions of the ratings themselves), we conclude that the presented rankings correlate strongly with the observed helpfulness votes.

The final true ranks of the top 30 reviews[7] will be used as an axis for some of the plots in our analysis. Anecdotally, we believe that these top 30 reviews are reasonably ordered by review quality, and we note that Amazon has a large commercial incentive to provide good rankings. Furthermore, we only consider the relative ordering of the top 30 results, and the top of the ranking has seen substantial attention from the users. But even if the final rankings do not reflect a perfect ordering by review quality, the voting patterns we identify below reveal strong dependencies that illuminate how users cast votes.

3. DO USER VOTES REVEAL THE CARDINAL QUALITY OF A REVIEW?

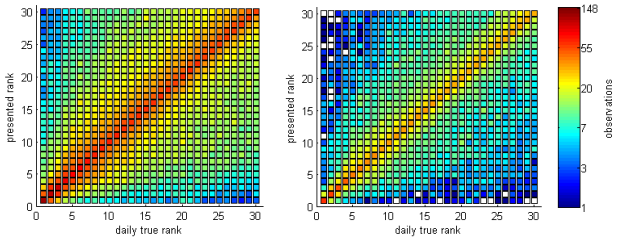
When users are asked to rate content, the final goal is to reach some assessment of the quality of this content. Throughout the following sections, we empirically investi-

[4]This is also the reason why we can have more than 29750 (595 products times 50) reviews in total, since we include all reviews that have ever been in the top 50 into the total, even if we do not see them in all snapshots.

[5]A natural concern could be that voting stops shortly after product's introduction and this results in skewed averages when using the whole length of the data.

[6]Reviews with the same number of helpful and unhelpful votes are considered tied; these largely occur between reviews with 0 or 1 total votes, of which we observe many in our dataset. Note that if we did not consider such reviews tied at the same rank, then rankings do not stabilize since Amazon appears to continue to reorder such reviews with the same (low) number of positive and negative votes.

[7]On our figures we use only the top 30 ranks out of 50 collected positions due to data sparsity.



(a) All (b) Late

Figure 2: Number of times (color in log scale, with red being high) a review with particular daily true rank (x-axis) was presented at a particular rank (y-axis). Left plot (*All*) counts over all data, right plot (*Late*) only over the last 30 days.

gate users' voting behavior when answering the question "Was this review helpful to you?", and how these votes relate to quality. We start by considering a natural and simple model of voting behavior, which was assumed in prior work [18, 4].

Cardinal Voting Model. The most natural model of how people vote is that they simply answer the question "Was this review helpful to you?" in an independent and objective way. Formally, this can be modeled as each different user u having a certain probability p_r^u of clicking "yes" for any particular review r. Over any distribution of users, the observed votes under this model follow a Bernoulli distribution, where $p_r = E[p_r^u]$ is the probability of observing a "yes" vote on review r, where the expectation is taken over the distribution of users:

$$P(yes|p_r) = p_r. \tag{1}$$

Under this model, p_r directly reflects the expected quality of review r in a cardinal manner, since p_r becomes synonymous with quality. Furthermore, estimating p_r for each review r can simply be done by using the observed fraction of "yes" and "no" votes, which is the maximum likelihood estimator:

$$p_r = \frac{\text{number of yes votes on } r}{\text{total number of votes on } r}. \tag{2}$$

We call this model of voting behavior the *Cardinal Voting Model (CVM)*. Note that this model, as well as the other models in this section, only describe the *polarity* of a vote, but not the decision of *whether* to cast a vote at all or not – this participation decision is studied in Section 5.

How accurate is this model of voting polarity and is it supported by our empirical data? Choosing an appropriate test to investigate the merit of the Cardinal Voting Model is somewhat subtle, and a number of obvious tests turn out to be flawed. For example, a natural test would be splitting the data into two cases: one includes all the instances where a review is presented above its final rank and the other the rest. Then we could use e.g. a paired Student t-test with null hypothesis that votes in both cases originate from the same distribution. If voting is based purely on a review's inherent characteristics alone — as in the CVM model — then being displayed above or below one's 'correct' rank (as given, e.g., by the final converged rankings) should not change the vote's polarity. However, this test introduces a bias: we only get a paired sample when we observe instances in both bins. For example, such a test would exclude all samples where currently over-ranked reviews get even more positive votes (and thus never become under-ranked), while including over-ranked reviews that do obtain negative votes (which would falsely support a hypothesis where users vote to fix the ranking).

In order to avoid flawed tests where the split of samples into cases to study voting patterns depends on the votes themselves, we use the following likelihood-ratio test that uses the Cardinal Voting Model as the null hypothesis. The test will identify whether presentation effects — even in their simplest form — can significantly better explain users' voting decisions.

Extended Cardinal Voting Model. Consider a simple extension of the CVM, where the probability with which a user votes "yes" or "no" on a particular review r not only depends on its inherent quality (as in the CVM), but also on the position where this review was presented in the ranked list. Specifically, the probability of a "yes" vote on a review with inherent quality parameter q_r and presented rank σ^{pres} is given by the following logistic model:

$$P(yes|q_r, \beta, \sigma^{pres}) = logit^{-1}(q_r + \beta\, \sigma^{pres}). \tag{3}$$

Note that the extended CVM model has one free parameter q_r for each review and a globally shared parameter β that models the influence of the presented rank σ^{pres} (which is observed). We estimate these parameters using maximum likelihood. This maximum-likelihood objective is convex, which means that the associated optimization problem can be solve globally optimally.

Note that the CVM is a special case of the extended CVM model with $\beta = 0$. The extended CVM with $\beta = 0$ merely parameterizes the CVM in terms of a quality parameter q_r, which is bijectively linked to the p_r parameter of the CVM through $p_r = logit^{-1}(q_r)$. Furthermore, maximum likelihood estimation in the extended CVM with $\beta = 0$ leads to exactly the same estimate of p_r as in Equation (2).

Testing the Influence of Presentation. The nested structure of the CVM and the extended CVM enables us to perform a likelihood ratio test that has the CVM model as its null hypothesis. Our test compares the likelihood of the observed data under the null hypothesis (i.e. $\beta = 0$) with the likelihood of the unrestricted model. If the improvement in likelihood is sufficiently large (*i.e.*, larger than one would expect from simply having more parameters to optimize over), then the likelihood ratio test rejects the null hypothesis.

The log-likelihood of the CVM ($\beta = 0$) model is -59267.78. The log-likelihood of the Extended CVM model is -57861.41. Both models are estimated from 136009 datapoints with 233512 total votes. The critical value according to the χ^2 statistics for one degree-of-freedom at the 95% confidence level is 3.84, which is much smaller than the observed difference in log-likelihoods of 1406.37. We can therefore reject the CVM model in favor of the extended CVM model with high confidence ($p < 0.001$). Clearly, users do not give independent assessments of the review quality. Simply presenting the review in a different position has a substantial effect on their voting behavior, which means that we cannot take the observed ratios of "yes" and "no" votes as a cardinal measure of quality as assumed in the CVM model.

In the rest of the paper we will explore improved models for how users make voting decisions. As a first step toward such models, let's look the the estimated β of the

Extended CVM model, which is $\beta = 0.0722$. Somewhat surprisingly, this value is positive, indicating that users are more likely to vote "yes" (as opposed to "no") if the review is presented *lower down in the ranking.* This stands in stark contrast to other settings where endorsements are used to rank items, especially Web Search where clicks are used as endorsements. Positive endorsements in Web Search follow a strong rich-get-richer pattern, where a result gets more positive endorsements (*i.e.*, clicks) the higher it is presented in the ranking [7]. While the types of endorsement (e.g., explicit positive and negative votes vs. positive clicks only) and the timing of the endorsement (e.g., before or after consuming the content) are different in the two settings, it is nevertheless evident that the two settings require very different machine learning methods for aggregating endorsements into an optimal ranking.

4. HOW DOES CONTEXT RELATE TO VOTING POLARITY?

The previous experiment showed that the rank at which a review is presented is correlated with the polarity of the vote users cast. Is there a plausible model of the user's decision process that could lead to this bias in user behavior? One can conjecture a large number of factors that causally influence a user's decision, ranging from position itself having causal effect, all the way to biases involving the time of day or week[8]. It is also conceivable that the history of votes so far also changes the user's perception of a review itself, as in herding phenomena [12]. Most promising, however, we conjecture that the *context* — the quality of surrounding reviews — might lead a user to change her opinion about the helpfulness of a given review. In particular, we explore in the following whether voting polarity shows any dependence on the degree to which a review is "misordered" relative to its context.

4.1 Statistical Analysis

We conducted a statistical analysis to see if voting polarity depended on misorderings in the ranking. To provide the tightest amount of control against confounding factors, we focus the statistical analysis on the voting behavior in the top three positions of the ranking — a more global analysis follows in the subsequent subsections. Since Amazon by default presents three reviews, the choice of three is natural. Let r_1, r_2 and r_3 be the best three reviews for a given product as determined by their *final ranks* (see Section 2.2). Table 1 compares the polarity of the votes on r_1 and r_2 under two different conditions, namely when the reviews were presented in the order $r_1 - r_2 - r_3$ vs. the order $r_2 - r_1 - r_3$. The average polarity is computed for each product and each condition separately (using only votes from those snapshots where the top three reviews appear in the desired locations) and the table shows macro averages over products. Note that the set of three results is the same under both rankings, and the key difference is the switch in ordering between r_1 and r_2.

Table 1 shows that the polarity of votes on r_1 is more positive in the swapped condition, while the polarity of votes on r_2 is more negative. Both differences are statistically significant according to a two-tailed paired Student t-test (a pair for each product) with $p < 0.05$, as shown in the last row of the table. This is in agreement with a model of user behavior where users cast their vote to "fix" the perceived misordering in the ranking by upvoting a review that is ranked too low, and downvoting a review that is ranked too high in relation to its context. Note that this is opposite to the biases identified for clicking behavior in web search that were observed in an analogous experiment [7].

Ordering	review r_1	review r_2
$r_1 - r_2 - r_3$	0.912	0.881
$r_2 - r_1 - r_3$	0.946	0.811
p-value	0.0395	0.0183

Table 1: Voting polarity when the same reviews are presented in different orders.

After this microscopic study of the first three positions (where a hypothesis test showed a significant depencence between review ordering and voting polarity), we now perform a macroscopic exploratory analysis of whether misorderings also correlate with observed changes in polarity over the whole ranking. To do this, we consider two different measures of how misordered a ranking is, which we call "global context" and "local context".

4.2 Exploratory Analysis: Global Context

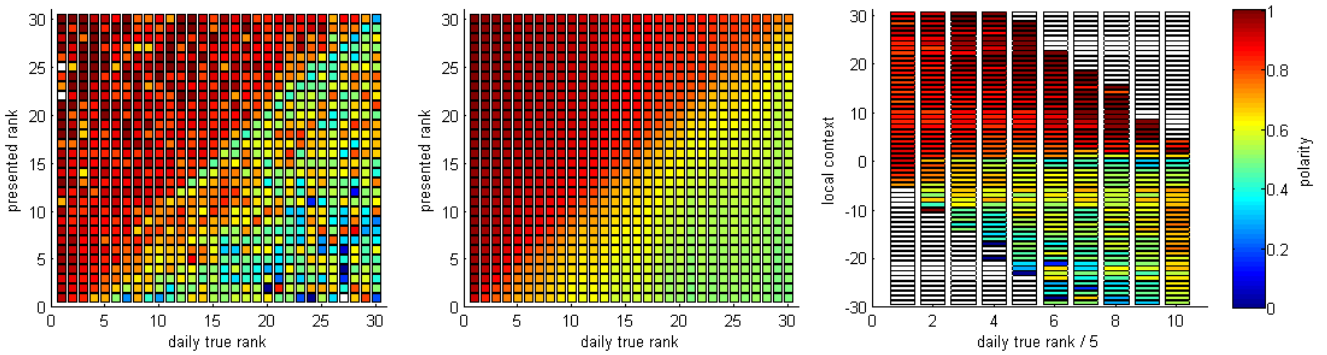
Our first measure of how misranked a particular review is relates the current position of a review to where it "should be" in a globally sorted ranking. To make this more precise, we define a quantity which we call the *daily true rank* of a review as follows.

Since new reviews for any given product might appear during the course of our data collection, we need a way to compute the "true rank" of a review — the "correct" rank that review "ought" to be displayed at — among the reviews present on any particular day[9]. To this effect, we disregard reviews which were not yet published by that day, and sort the already existing reviews by their final true rank. This results in the daily true rank of a review. We call a review overranked if it is presented above its daily true rank, and underranked if it is presented below.

Figure 3(a) displays the average polarity of the votes (color from negative in blue to positive in red) as a function of daily true rank on the x-axis and presented rank on the y-axis. The figure shows that the more a review is presented above the rank it deserves, the more negative the polarity of the votes is. On the other hand, when a review is presented too low relative to its deserved rank, the polarity of the votes is more positive. These observations hold across a wide range of presented ranks and daily true ranks. This can be interpreted as users upvoting reviews that are rated 'too low', and downvoting reviews that are rated 'too high'. Note that we carefully define our notion of converged ranks, using rankings collected 4 months after the end of our daily snapshot collection to mitigate self-fulfilling prophecies in terms of using the same votes to understand behavior as to determine what is too low or high, as described in Section 2.2. Furthermore, by fixing the daily true rank (each

[8]For example, the user population visiting the site primarily during weekends might be in a better overall mood and thus vote more positively.

[9]Note that this can be different from the set of all reviews present on the last day



(a) global context, data (b) global context, model (c) local context, data

Figure 3: Vote polarity: daily true rank (x-axis, with bin widths of 5 for local context plot), context (y-axis, presented rank in the case of global context, positive values meaning superior in the case of local context) and average vote polarity (color, red means higher ratio of positive votes).

one has its own bin on x-axis) we avoid a possible bias of it correlating with the likelihood of being over-ranked (e.g., the best review can never be over-ranked) and skewing the averages (*i.e.*, positive votes from good reviews contributing mainly to under-ranked polarity average).

Voting Polarity over Time. Because we collected our data over a long timespan (a few months) there might be differences in the voting behavior between early and late periods. To explore this, we separately plot analogues of Figure 3(a) using only data from the first and the last 30 days of our collection period, respectively. The resulting plot are shown in Figures 4(a) and 4(b). While the plots are more noisy due to the smaller datasets, the observed patterns are remarkably stable over time.

4.3 Exploratory Analysis: Local Context

Using global context to model a user's perception of how misordered the ranking is assumes that the user has a global understanding of the ranking. This clearly can only be true in an approximate sense. More likely, a user bases her voting decision on a more local view of misordering. Since Amazon displays reviews as a ranked list, a user will see any particular review in the context of the other reviews surrounding it at that time. We refer to the reviews presented immediately above and below a particular review as the *local context* of the review. For our data analysis, we define the local context as the 3 reviews appearing immediately above and below a review at any given time unless otherwise specified[10].

Local superiority and inferiority. We use the following measure to capture how the quality of a review relates to the 6 reviews in its local context: we take the average of the daily true ranks of these 6 reviews and subtract the daily true rank of the review under consideration. When this difference is positive, we say the review is *locally superior* (of higher quality compared to its surrounding reviews), and *locally inferior* when the difference is negative.

Figure 3(c) demonstrates that local context does indeed correlate with the polarity of votes received by a review. The x-axis ranges over values of the daily true rank (each bin is 5 ranks wide), and the y-axis measures the relative quality, where negative values correspond to local inferiority and positive to local superiority. The color of a point (x, y) is the polarity of votes received by a review when it has daily true rank x and local context y. If local context did not correlate with voting patters, there should be no gradient in color along the y-axis. However, Figure 3(c) displays a noticeable increase in the fraction of positive votes when a review is locally superior, and vice versa when the review is locally inferior.

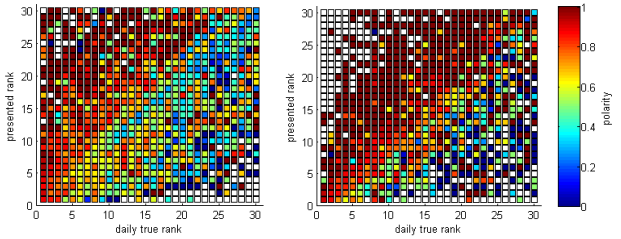
Note that the effect of global and local context on voting polarity closely resemble each other: a review receives more positive votes when it is under-ranked (a global measure) as well as when it is locally superior, and fewer positive votes when it is over-ranked as well as when it is locally inferior. That is, the interpretation that users vote to correct misorderings in the ranking is not sensitive to our particular measure for misordering, and in fact leads to similar qualitative observations for two different measures of misordered rankings.

4.4 A Model of Voting Polarity

In this section, we abstract the observed voting polarity patterns into a general model. The reason for this is twofold. First, such models of voting polarity are needed for designing content-ranking algorithms that make optimal use of the votes they elicit. Second, a formal model can be tested and verified also in other content-rating settings.

We model voting polarity using a traditional form of logistic regression (with two variables capturing insights gained in the previous sections). Each review r has an inherent quality q_r that is unknown. In addition to inherent quality, our model considers the context δ^{ctxt} of a review at the time a user casts a vote. A positive δ^{ctxt} means that review r is better than its context, a negative δ^{ctxt} means that it is worse. The strength with which δ^{ctxt} influences polarity is

[10]We ignore presentation-induced anomalies such as those created by page breaks, *i.e.*, the fact that reviews near the top or bottom of a page are not visually surrounded by their local context on both side in the same pageview. Also, we verified that our specific choice of the number 3, as well as considering reviews appearing only above or below a review does not alter our results (since it is conceivable that reviews appearing immediately before a review would influence the vote cast on that review more than those that are read after it); the pattern of the fraction of positive votes received as a function of local context appears to be quite robust to the specifics of exactly how the local context is defined.



(a) Early (b) Late

Figure 4: Vote polarity: daily true rank (x-axis), presented rank (y-axis) and vote polarity (color, red being positive). Only using the first 30 days (left) and last 30 days (right) of data.

captured by the parameter β:

$$P(yes|q_r, \beta, \delta^{ctxt}) = logit^{-1}(q_r + \beta\, \delta^{ctxt}). \quad (4)$$

Both global context (*i.e.*, the difference between presented rank and daily true rank) and local context as defined above can be used in approximations of context. However, it is unreasonable to assume a linear relationship between δ^{ctxt} and these rank-based measures. Instead, we use the following transfer function that de-emphasizes the impact of large rank differences. Let δ^{ctxt}_{rank} be either the local or the global context in terms of rank, then

$$\delta^{ctxt} = sign\left(\delta^{ctxt}_{rank}\right) log\left(1 + \left|\delta^{ctxt}_{rank}\right|\right). \quad (5)$$

Using this transfer function and global context as a proxy for δ^{ctxt}, the fitted model has a log-likelihood of -57051 with parameter $\beta = 0.415$ (under-ranked reviews have more positive δ^{ctxt} which in turn means more positive polarity due to a positive β). Figure 3(b) plots the fitted model, which can be compared directly to Figure 3(a). In particular, Figure 3(b) is produced by replaying the observed rankings that produced Figure 3(a) and then, for each cell, averaging the predicted probabilities of the model instead of the observed vote polarities. Overall, the model captures the key trends in the data, including a decrease in voting polarity with rank on the diagonal, and the increase in voting polarity for reviews that are ranked too low.

5. HOW DOES CONTEXT RELATE TO PARTICIPATION?

We have so far only explored user voting behavior in terms of *polarity* (*i.e.*, how the users cast "yes" vs. "no" votes). A second, equally important dimension is participation — *when* does a review receive votes? A first guess would be that the number of votes a review receives depends largely on its presented rank, corresponding to an attention bias whereby more users read, and therefore vote on, reviews that are displayed at higher rather than lower ranks. However, this is not the entire story — participation also depends on context, as we show below.

Ideally, participation would be measured as the ratio of the total number (positive plus negative) of votes on a review to the number of views it receives. However, since our data does not include information about pageviews, we make the assumption that there is a constant number of pageviews (for each position) in each period between snapshots (note that the actual number of such pageviews does not matter since we are only interested in relative comparisons). An additional issue is sparsity of data — most reviews receive no or one vote on most days. To deal with these issues, we measure participation using the following statistic. Consider a particular bin (x, y) defined by some value of the x and y variables (daily true rank and local/global context). As a measure of participation, we use the ratio of the number of intervals (between two consecutive snapshots) where at least one new vote was cast on reviews in this bin to the total number of observed intervals in the bin.

5.1 Statistical Analysis

To investigate whether there is a dependence of participation on context similar to polarity, let us start by considering the same experiment setup as in Section 4.1. Table 2 shows participation at rank 1 and rank 2 conditioned on the ordering of the top two results. When r_1 is correctly ordered before r_2, there is a strong decay in participation, as one would expect from an attention bias (see [7] for similar attention biases in web search). Once r_1 gets misordered into position 2, however, voting participation on position 2 significantly (two-tailed paired Student t-test, $p < 0.05$) increases compared to the correct ordering. It appears that users are motivated to participate once they see that the ranking is misordered. Voting participation on rank 1 does not change significantly. A plausible explanation is that users typically do not go "back" up the ranking to vote once they have realized that there was a misordering.

	rank 1	rank 2
$r_1 - r_2 - r_3$	4.34	2.33
$r_2 - r_1 - r_3$	3.71	4.19
p-value	0.395	0.004

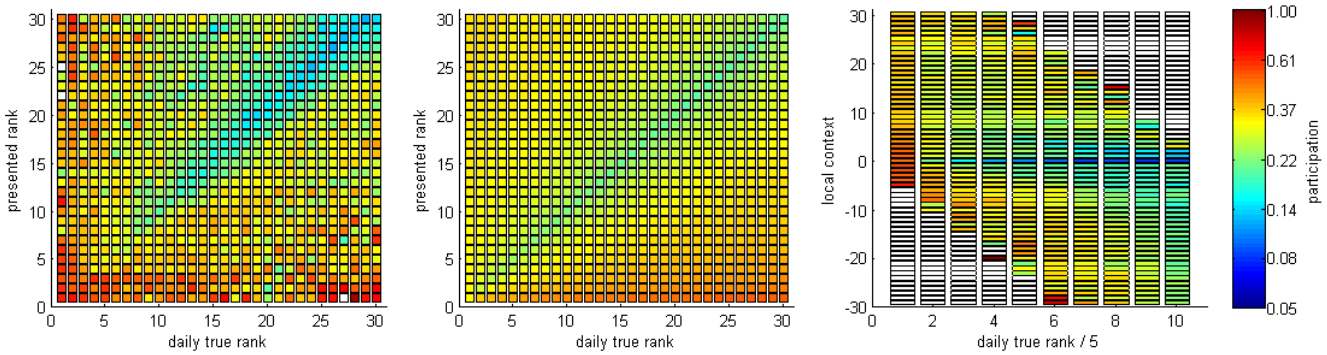
Table 2: Participation when the same reviews are presented in different orders.

Similar to our argument for polarity, we now explore in how far this microscopic finding about participation extends more macroscopically.

5.2 Exploratory Analysis: Context

We first investigate how participation varies as a function of global context. Figure 5(a) plots participation (from low participation in blue to high participation in red) as a function of presented rank and daily true rank. As expected, there is an attention bias. Amongst reviews that are correctly ranked (*i.e.*, presented at their daily true rank), reviews with higher ranks are voted on more often (note that the color scale is logarithmic). This can be observed from the diagonal elements in Figure 5(a). Furthermore, note that there is a special attention bias for the top three presented ranks, which receive much larger participation. This is because Amazon presents the top 3 reviews on the product page, while an additional click is required to display more of the ranking.

Beyond attention bias, Figure 5(a) shows that context affects participation as well — reviews receive more votes when they are incorrectly ranked. Comparing the off-diagonal elements in Figure 5(a) with the diagonal ones, we see that



(a) global context, data (b) global context, model (c) local context, data

Figure 5: Participation: daily true rank (x-axis, using bin widths of 5 for local context plot), context (y-axis, presented rank in the case of global context, positive values meaning superior in the case of local xt) and average participation (color, log scale, red being high).

reviews get voted on more often when they are not in their correct position. The upper triangular portion of Figure 5(a) demonstrates the same effect that was already observed in previous subsection's experiment — a good review that is ranked too low receives more votes. Unlike in the statistical analysis of the top three positions, however, the lower triangular portion of Figure 5(a) shows that participation at lower presented ranks also increases when a review is presented too high. A plausible explanation is that for reviews at lower presented ranks, the user has already formed a reliable expectation about the review quality of the next review, upon which the user can judge misordering.

Similar trends also emerge in the analogous Figure 5(c) for local context. Reviews that are locally superior or inferior (non-zero bins on y-axis) get voted on more often.

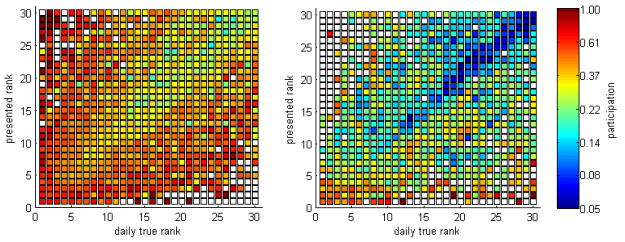
Voting Participation over Time. Similarly to polarity (Section 4.2), we observe temporal consistency in voting participation. Both Figure 6(a) and Figure 6(b) are analogous to Figure 5(a), but are restricted to the first 30 days and the last 30 days of the data collection period, respectively. Both plots show the same general v-shaped pattern, indicating that participation follows a stable pattern over time.

5.3 A Model of Participation

The following proposes a model of participation that is analogous to the one derived in Section 4.4 for polarity. For each position p, we model the "normal" amount of attention a review at this rank gets using the parameter z_p. Variable δ^{ctxt} is the context of review r as defined for polarity, and we use the same transfer function from Equation 5 to connect δ^{ctxt} to the rank-based measures of global and local context. However, since participation is symmetric in δ^{ctxt}, we use its absolute value. The parameter α models the influence of context:

$$P(vote|z_p, \alpha, \delta^{ctxt}) = logit^{-1}\left(z_p + \alpha \left|\delta^{ctxt}\right|\right). \quad (6)$$

We fit this participation model directly to the observed participation frequencies from Figure 3(a) using maximum likelihood, where each cell receives the same weight in the maximum likelihood estimate. We do this because the number of observations in the first few presented ranks otherwise dominates the likelihood and biases the parameters towards a fit to only those cells. We also smooth the parameters z_p to lie on a curve $z_p = \gamma_0 + \gamma_1/p^{\gamma_2}$, where γ_0, γ_1, and γ_2 are fitted parameters. This step is added to produce smoother estimates of the z_p despite data sparsity at low ranks, but it is not essential for capturing the general trends.



(a) Early (b) Late

Figure 6: Participation: daily true rank (x-axis), presented rank (y-axis) and participation on log scale (color, red being high). Only using the first or last 30 days of data.

The resulting model is plotted in Figure 5(b), and it is analogous to the plot of the observed data in Figure 5(a). The parameter α of the fitted model is 0.228. This means that the more misranked a review is according to $|\delta^{ctxt}|$, the higher the participation. While the model underestimates participation at the top presented ranks (which is due to the equal weighting of the cells during maximum likelihood estimation), the plot overall resembles the patterns in the observed data.

6. SUMMARY AND DISCUSSION

In this paper, we analyzed user voting patterns on Amazon reviews, over time and as a function of the context in which the review was rated by the user. Using a dataset of daily snapshots of reviews and their ratings, we observed that voting *polarity* (*i.e.*, whether to assign a positive or negative helpfulness vote to a review) as well as *participation* (*i.e.*, whether to vote at all) depends not only on the inherent quality of a review, but also on the context in which it is presented at the time of voting. In particular, we provided evidence that the observed connection between context and

voting behavior cannot be captured by a cardinal voting model (Section 3), where users make absolute and independent judgments about helpfulness.

As an alternative to the cardinal voting model, we proposed models of voting that incorporate context in addition to a review's inherent quality, which we show to provide a much closer fit to the observed data. Notably, we find that voting polarity becomes more positive/negative, if a review is better/worse than its context. Furthermore, we find that voting participation generally increases if a review is misranked. These patterns are substantially different from other setting where endorsements are used to rank items, most prominently Web Search, which means that methods for learning rankings in one setting cannot be naively transferred to another.

Future Work and Applications. The insights gained in this paper have many practical applications. For example, can the ranking algorithms be improved, if we know that users do not vote purely according to inherent review qualities? In our future work we plan to explore how one can learn rankings faster, given this improved understanding of users' voting behavior. Another implication of this work is an interaction design question, namely, how should we phrase the feedback question. Considering that users apparently do not answer all questions as they are asked (*i.e.*, Amazon asks for pure quality feedback but gets also some context dependent answers), are there other feedback questions that provide higher-quality feedback or improve participation?

6.1 Limitations

As we have already pointed out throughout the paper, the study design and analysis have some limitations which we further discuss below. These provide several interesting directions for further work.

(i) *Observational Data.* Our analysis, which is based on observational rather than experimental data, comes with the limitations typically associated with deriving models from observational data. While non-uniformity in the presented ranks provided natural experiments which allowed us to investigate correlations, only experiments with controlled interventions can ultimately validate the causal reliability of our conclusions and models.

(ii) *Snapshot Granularity.* Recording one snapshot per day is too infrequent to capture all the voting events that occur on Amazon. In particular, there are reviews that receive more than a hundred votes in a single day, and rankings can also change very fast. We collected a separate dataset that contained snapshots at intervals of about one hour collected over the course of one day; even here there were products where review rankings changed between snapshots. However, our observations on this dataset showed that rankings do not change drastically (e.g., often only reviews with the same number of votes swapping places), so that using daily snapshots does provide an acceptable approximation for studying voting patterns on typical reviews on Amazon. Nevertheless, studying a more fine-grained dataset can potentially lead to new insights on very frequently read reviews.

(iii) *No Independent Assessments of True Quality.* Ideally, we would have liked to use independent assessments to determine the true quality (*i.e.*, true rank) of each review where we use this quantity in our analysis. However, this was infeasible since even with access to grading resources, it would have been unclear how to judge which reviews truly are helpful to the (unknown) user population. A possible concern with an analysis that uses final ranks as a proxy for true quality is that the data that we use to analyze voting behavior also contributes to defining the true qualities (namely the final ranks): for example, a review with a given final rank k might have been ranked lower, *i.e.*, been underranked, at some earlier point in our dataset, and then received an above-average share (relative to the entire set of votes that contribute to the final rank) of positive votes after this time to climb up to its final rank. This can lead to a self-fulfilling prophecy where 'underranked' reviews receive an above-average share of positive votes, and 'overranked' reviews a below-average share. The following points help mitigate concerns regarding such self-fulfilling prophecies: (a) We compute the final true ranks based on a snapshot recorded 4 months after the last sample in the data we use for our analysis of voting patterns. That is, our final ranks also include a large number of votes cast over a long interval that does not overlap with the interval corresponding to the votes in our empirical analysis (see Section 2.2); (b) The statistical analysis in Section 3 does not make use of final ranks at all, but still shows that voting is not based purely on inherent review quality.

(iv) *Aggregate vs. Individual Voting Models.* Our models do not claim to describe the actions of any individual user, but merely the aggregate behavior of a user population. For example, we cannot ask whether each individual user modifies polarity based on rank (even though that is a plausible conjecture), but only observe that the population of users displays this aggregate behavior. The same aggregate behavior, however, could also be explained by other factors, such as heterogeneity in user populations (different types of users may have different baseline polarities and may explore rankings to different depths). Investigating the effect of such alternate hypotheses is an interesting open direction.

(v) *Macroscopic Participation and Polarity.* Our model for participation predicts that a review displayed at its correct rank should receive fewer votes. This means once rankings converge to the right ordering of reviews, participation should globally decrease. This is indeed observed in the data. Figure 2(b) shows that rankings stabilize after a few months, and Figure 1 shows that voting also decreases at this time. Regarding voting polarity, Figure 7 shows that the average vote polarity becomes more positive with time. This is to be expected. Since converged rankings have 'good' reviews at the top (with high attention bias), these reviews have to maintain a high fraction of yes votes (or regain that ratio once their rank has dropped). Trends in participation and polarity could also have other causes, however. For example, participation could be explained by a product becoming outdated and consequently fewer users reading and voting on reviews. Similarly, an alternate explanation for polarity is that users continue to give positive votes to reviews they like even when rankings are accurate, but do not downvote reviews unless necessary to correct the ranking.

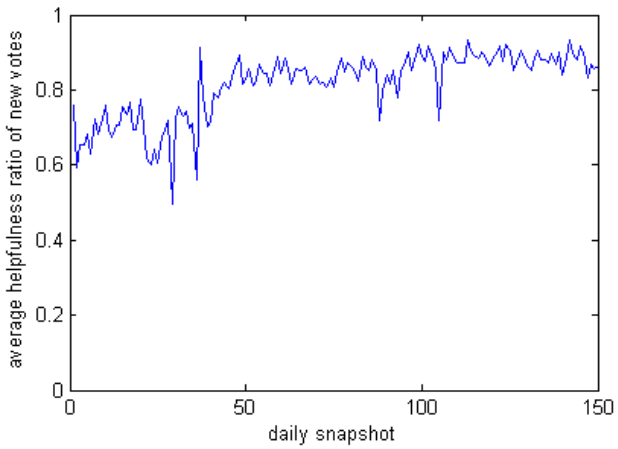


Figure 7: Average ratio of positive to total votes (y-axis) over time (x-axis in days).

(vi) *Perception versus Action.* A final intriguing question that we do not address is the mechanism by which context affects voting: does a review's context cause a user to actually change her perception of its helpfulness, or merely her action? One possibility is that a user may truly value a review less if it is presented next to an even better review. A different possibility is that each user might have some cardinal quality capturing each review's helpfulness, and vote in order to 'correct' the current vote ratio (of current "yes" to current total votes) towards her opinion of the review's quality. Preliminary explorations indicate that both effects (modified perception and modified action) might be supported by the data. An experimental study, or analysis of a more detailed dataset, could lead to clearer conclusions on this question.

7. RELATED WORK

Presenting the best content to viewers is an essential component to the value as well as commercial success of on content-centric websites, whether measured by the number of unique visitors to the site or its total sales revenues. Consequently, there has been plenty of research on various aspects of ratings and the quality of online content.

The majority of this literature has focused on the natural problem of machine-based methods for inferring and predicting content quality, since the large volume of user-generated content rules out the possibility of trusted human editors rating and ranking all contributions to a site. [8, 18, 16, 6, 10, 13, 5, 15] all address the fundamental question of what features of a contribution can help accurately predict its quality, where the gold standard for quality is based either on ratings by independent human raters [9, 15], or the actual received upvotes and downvotes on the website [8, 18, 16, 10, 15]. A number of features are found to be predictors of (appropriate notions of) quality in various settings, ranging from textual content, review length, star rating and product category for Amazon reviews [8, 18, 13], to comment sentiment in Youtube [16], to the topic and payment amount on online Q&A sites such as Google Answers [5]. Later work uses increasingly sophisticated features such as social context [1, 11] in addition to the textual content of a contribution to improve prediction accuracy. The key difference between this literature and our work is that this literature focuses on the problem of inferring the quality of a contribution, typically using classification or regression-based approaches, whereas we are interested in the behavior of the *raters* themselves.

There is a much smaller literature on ratings of online content that takes the perspective of understanding or modeling rater behavior. [3] study helpfulness ratings on Amazon in the framework of *opinion evaluation*, and find that social factors influence users' ratings of opinions (here, reviews). Specifically, factors such as how closely an opinion (expressed by a review) agrees with other opinions on the same issue (*i.e.*, the product being reviewed) — as measured by the star rating associated with the review text in relation to other reviews' star ratings — influence users' ratings of each opinion. More recently, there has been experimental work on *herding* effects on user ratings of online reviews [17, 12, 14], investigating how a user's awareness of previous votes on a review impacts his own voting decision (e.g. users being more likely to cast a positive vote after seeing previous positive votes compared to the case of not being able to observe the previous votes). Finally, [9] notes biases in user voting behavior as causing a deviation between gold standard ratings from independent human raters and helpfulness votes. In particular, the imbalance bias mentioned in [9] is potentially closely related to the issue of voting participation, since it suggests that users might preferentially choose to vote when they have a positive rather than a negative opinion of a review's helpfulness. However, [9] only cites user behavior as a possible cause for divergence between gold standard ratings of quality and aggregated helpfulness votes, and primarily focuses on the problem of identifying features to build an accurate classifier for separating high and low quality reviews.

Finally, there is a small literature [2, 4] that looks at theoretical questions regarding ranking and incentivizing high-quality contributions respectively assuming simple quality-based Bernoulli models of voting behavior. Our empirical analysis of user voting patterns can potentially supply richer, more realistic models of rating behavior upon which to base such theoretical studies on algorithms and mechanisms for eliciting and identifying high-quality online content.

8. ACKNOWLEDGEMENTS

This research was funded in part by NSF Awards IIS-1217686 and IIS-1247696, and the Cornell-Technion Research Fund.

9. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194, New York, NY, USA, 2008. ACM.

[2] G. Askalidis and G. Stoddard. A theoretical analysis of crowdsourced content curation. In *The 3rd Workshop on Social Computing and User Generated Content*, 2013.

[3] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. 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*, WWW '09, pages 141–150, New York, NY, USA, 2009. ACM.

[4] A. Ghosh and P. McAfee. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 137–146, New York, NY, USA, 2011. ACM.

[5] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online Q&A sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 865–874, New York, NY, USA, 2008. ACM.

[6] N. Jindal and B. Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1189–1190, New York, NY, USA, 2007. ACM.

[7] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2), April 2007.

[8] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 423–430, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[9] J. Liu, Y. Cao, C. Y. Lin, Y. Huang, and M. Zhou. 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 (EMNLP-CoNLL)*, pages 334–342, 2007.

[10] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 443–452, 2008.

[11] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 691–700, New York, NY, USA, 2010. ACM.

[12] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[13] S. M. Mudambi and D. Schuff. What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Q.*, 34(1):185–200, Mar. 2010.

[14] M. Nakayama and Y. Wan. An exploratory study: "Blind-testing" consumers how they rate helpfulness of online reviews. In *Conf-IRM*, 2012.

[15] S. Sen, F. M. Harper, A. LaPitz, and J. Riedl. The quest for quality tags. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 361–370, New York, NY, USA, 2007. ACM.

[16] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro. How useful are your comments?: Analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 891–900, New York, NY, USA, 2010. ACM.

[17] F. Wu and B. Huberman. How public opinion forms. In C. Papadimitriou and S. Zhang, editors, *Internet and Network Economics*, volume 5385 of *Lecture Notes in Computer Science*, pages 334–341. Springer Berlin Heidelberg, 2008.

[18] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 51–57, New York, NY, USA, 2006. ACM.