Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

To Buy or to Read: How a Platform Shapes Reviewing Behavior

Edward Newell[1], Stefan Dimitrov[1], Andrew Piper[2], Derek Ruths[1]
McGill University 845 Sherbrooke Street West, Montreal, Quebec, Canada H3A 0G4
[1]School of Computer Science, [2]Department of Languages, Literatures, and Cultures
{edward.newell, stefan.dimitrov}@mail.mcgill.ca
{derek.ruths, andrew.piper}@mcgill.ca

Abstract

We explore how platforms influence user-generated content by comparing reviews made on the retail platform Amazon, with those on the (non-retail) community platform Goodreads. We find the retail setting gives rise to shorter, more declamatory, and persuasive reviews, while the non-retail community generates longer, more reflective, tentative reviews with more diverse punctuation. These differences are pronounced enough to enable automatic inference of the platform from which reviews were taken with over 90% F_1. Significant differences in star-ratings appear to parallel differences in review content. Both platforms allow users to give feedback on reviews. Experiments show that a subtle difference in the review feedback features influences review-promotion behavior, which may in part explain the differences in review content. Our results show that the context and design of a platform has a strong but subtle effect on how users write and engage with content.

Introduction

Many websites host user-generated content, which can generate valuable original content and encourage deeper user engagement. How platform design influences user-generated content remains an open question. We investigate the case of book reviews, and compare reviews made on the retail platform Amazon, with those made on the (non-retail) enthusiast community platform Goodreads. Since one platform is retail-oriented, we expect users visit the platforms for different reasons, which might lead to different review writing practices.

Comparing reviews for the same book helps to control for the topic of the review. By selecting the New York Times Bestsellers, it also helps focus on sampling a common set of users: avid readers of best-selling English-language books. Reviews follow a consistent template — the review content, the star-rating, and the community evaluation (e.g., upvoting or liking) — which enables us to compare reviews according to each aspect. This has implications for both users and platform designers: understanding how reviews differ between platforms could help guide users toward more informative sources, and help uncover the platform design factors that elicit informative reviews.

We begin by showing that reviews from the platforms differ significantly in stylistic markers. We show that a classifier can distinguish ($F_1 > 90\%$) between reviews from each platform, making it plausible that these differences are perceivable by users. We find that the distributions of star-ratings also differ significantly, in a way that parallels the differences in content, and in a way that can be explained by the platforms' underlying purposes.

We then analyze how each platform lets users give feedback on reviews (akin to upvoting and downvoting). Using a controlled experiment, we show that a subtle difference in the feedback features on either platform leads people to promote different kinds of reviews. This demonstrates how small differences in the design of the platforms could give rise to different review-writing practices.

Despite soliciting reviews on identical objects (books), Goodreads and Amazon elicit different reviewing behaviors. These differences can be related to the purpose of each platform, suggesting that platform design influences review behavior in subtle ways.

Related Work

Product reviewing behavior. Mudambi and Schuff 2010 identified three factors which are related to review helpfulness: review extremity, review depth and product type. We note that, while they found extreme reviews were less helpful, their findings related to experience-based goods, while for books, another study found review extremity correlates with helpfulness (Forman, Ghose, and Wiesenfeld 2008). Review age has been found to inversely correlate with helpfulness (Otterbacher 2009). However, the default sorting of reviews according to existing votes can lead to a "winner circle" bias (Liu et al. 2007). Another study showed that 10%-15% of reviews highly resemble previously posted ones (Gilbert and Karahalios 2010), and that there are both amateur and professional review writers.

Platform influence on reviewing behavior. Prior work identifies three ways in which a platform can influence user-generated content: platform design, social conformity, and moderation.

Platform design factors include aesthetics, priming through Captchas, and explicit prompts in web forms (Sukumaran et al. 2011). Social conformity refers to the tendency for users to adapt their contributions to be more like pre-

existing ones (Sukumaran et al. 2011; Michael and Otterbacher 2014). Moderation is a means of enforcing standards on user-generated content, by explicitly flagging, grading, or filtering content. Amazon employs what has been called *pre-moderation* (Diakopoulos and Naaman 2011), whereby comments must be approved by site administrators before appearing. Both platforms use *post-moderation* (Diakopoulos and Naaman 2011), by enabling users to provide feedback on reviews.

These prior works highlight various mechanisms by which platforms might shape user-generated product reviews, which motivates the design of our study. It remains to be shown however, whether operational platforms in fact give rise to observable effects, which we investigate here.

Cross-platform studies. Dimitrov et al. 2015 also compare book reviews from Amazon and Goodreads. However, they only consider the biography genre; we expand this analysis to a much wider variety of genres. We show that previous findings with respect to star-ratings hold in the broader context, but also undertake an analysis of linguistic differences. We demonstrate that differences in review content are sufficient to enable automatic inference of the platform from which the reviews were taken, which establishes that these differences plausibly affect user experience. Dimitrov et al. offer an observational analysis, but we demonstrate a potential mechanism of influence. We experimentally demonstrate that subtle differences in the features enabling user-feedback on reviews leads to the promotion of different reviews.

Platforms and Dataset

Comparison of platform designs

At a superficial level, the book detail pages of the platforms are fairly similar—both show the book's cover prominently with a short description to its right, with users' reviews in a section below this. However, Amazon features more prominent retail-related elements, while Goodreads provides a single discrete link to buy the book on Amazon. Reviews on Goodreads show reviewers' profile pictures, unlike on Amazon, and on Amazon, users are allowed to make anonymous reviews. Amazon reviews include a title where users summarize their attitude towards the book. Both platforms let registered users write responses (comments) to existing reviews. By default, Amazon sorts reviews by helpfulness; Goodreads uses a proprietary sorting algorithm which considers the number of likes and other undisclosed features. We highlight an important difference in the mechanism by which users provide feedback on reviews: on Amazon users can flag a review as "helpful" or as "not helpful", on Goodreads users can "like" a review, but there is no corresponding "dislike" option.

Dataset

Between February and April 2015, we collected the 60 most recent reviews for 3,381 New York Times bestsellers[1] as listed from January 3, 2010 through January 3, 2015. After removing non-English reviews using langid (Lui and Baldwin 2012) (5,195 from Goodreads and 31 from Amazon), and eliminating duplicates, we obtained 189,329 Goodreads and 195,195 Amazon reviews.

[1] www.nytimes.com/best-sellers-books/

Dictionary	Amazon	Goodreads	Difference
Linguistic processes			
Swear words	0.04	0.07	-75.0
Numerals	0.34	0.52	-52.9
Exclamation marks	1.12	0.60	+46.4
Question marks	0.17	0.22	-29.4
Parentheses	0.22	0.33	-50.0
Colons	0.11	0.16	-45.4
Commas	3.37	4.05	-20.2
Dashes	1.06	1.35	-27.4
Other punctuation	0.25	0.35	-40.0
"you"	0.79	0.57	+27.8
Psychological processes			
Positive emotions	6.66	5.16	+22.5
Tentative language	2.21	2.54	-14.9
Certainty language	1.76	1.50	+14.8
Negative emotions	1.71	1.95	-14.0
Anger	0.53	0.69	-30.2

Table 1: Frequency of linguistic features (expressed as % of words belonging to each dictionary) in reviews, by platform. All statistics shown are significant with $p < 0.001$. Difference is relative to Amazon, reported as a percentage, with positive numbers indicating higher frequency on Amazon.

Analysis of content

We take as our first hypothesis:

H1: *The design and function of a platform influences the content of reviews.*

Discussion of buying experience

We begin by manually annotating 1000 random Amazon reviews, indicating whether they discussed the buying experience (purchase, delivery, shipping, or transaction), which we expected to be a distinct feature of Amazon reviews. Surprisingly, only 19 of the 1,000 manually annotated reviews mentioned the buying experience in any way. Thus, differences which we do find would have to be attributed to more subtle effects on how review authors write, rather than on the explicit tendency to focus on topics related to purchasing.

Linguistic analysis

We looked for subtler effects on writing style, using the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker, Francis, and Booth 2001). This tool identifies the prevalence of style markers, such as certain punctuation marks

and words from specific sets (dictionaries). For each platform, we concatenated all reviews for a given book into one *ensemble*, and subjected the ensembles to LIWC analysis (Table 1).

Amazon reviews use more exclamation marks and words connoting certainty (e.g., "unambigu*", "fundamentals", "perfect*", "always", and "guarant*"). Amazon reviews employ the pronoun "you" more frequently, and tend to use more positive language. In contrast, Goodreads reviews use more diverse punctuation, more tentative words (especially those of a colloquial nature, e.g. "lotsa", "dunno", "shaky", "kinda", etc.), and more balanced sentiment.

Amazon reviews appear to be oriented around a more persuasive, positive, declamatory style, as one might use to convince others to buy or not to buy a book. On the other hand, Goodreads reviews appear more oriented around community discussion, with expressive pauses, critical viewpoints, and colloquial hesitations, suggesting a greater degree of self-reflection and an effort to participate in a social community of readers.

Review Classification

We next sought to determine whether these stylistic differences could plausibly differentiate the platforms. We trained classifiers to infer the origin of review *ensembles*, which bundle all reviews from our sample for a given book on a given platform. Classification for ensembles was based on features at different levels of granularity (Table 2). Review and sentence length features were quantized using length intervals selected based on information gain. The LIWC dictionary features counted the prevalence of words from each of its 65 specialized dictionaries. The "literary features" comprised counts for 101 words related to literary analysis selected by a literature expert (e.g. "narrative", "point of view", "protagonist", etc.). A final set of features counted references to book titles and authors. Altogether there were 178 individual features (see Table 2).

The classifiers achieved 92% and 95% F_1 for fiction and non-fiction books respectively (based on 10% held-out data). We used 10-fold cross validation for hyperparameter tuning and kernel function selection for a support vector machine from Scikit-Learn (Pedregosa et al. 2011), with a radial basis kernel performing best.

Interestingly, an initial attempt to classify individual reviews (as opposed to ensembles) performed poorly (62% overall accuracy). This suggests that there is significant within-platform variability between reviews, preventing individual review classification, but that the different platforms influence review writing in ways that build up within ensembles. We expect individuals' user experience to be akin to the ensemble classification: over time, users will encounter many dozens of reviews, making small linguistic and stylistic differences pronounced.

Ablation testing (Table 2) shows that review length is the most discriminative feature, with the LIWC dictionaries second. Review length and vocabulary are both readily perceived by users, so might be reinforced by social conformity.

Feature set	# Features	Fiction	Non-Fiction
Review-level features			
Review length	6	0.92	0.86
Book title / author	4	0.63	0.61
Sentence-level features			
Sentence length	2	0.62	0.62
Word-level features			
LIWC dictionaries	65	0.84	0.77
Literary features	101	0.70	0. 64

Table 2: Feature sets used for classification of reviews. For each, the number of individual features in the set is shown, along with F1 scores on fiction and non-fiction review ensembles.

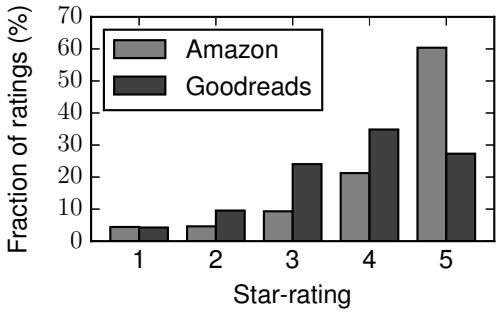

Figure 1: Distribution of ratings over all books on Amazon and Goodreads.

Analysis of Star-Ratings

We begin our analysis of star-ratings with the following hypothesis:

H2: *Ratings are more extreme in a commercial context (Amazon) than in a non-commercial one (Goodreads).*

In general, reviewers of both platforms prefer to give ratings of four or more stars (Figure 1), but we see marked differences between the ratings on each platform ($p < 0.001$). We define extreme ratings to be ratings of one or five stars. A two-proportion z-test shows a significantly higher proportion of extreme ratings on Amazon ($p < 0.001$). The more moderate ratings on Goodreads align with the finding that Goodreads reviews employ more balanced proportions of positive and negative emotions, and more tentative language.

Analysis of Review Promotion

Users on Amazon may flag reviews as "(un)helpful", while on Goodreads users can "like" reviews. We hypothesize that this subtle difference influences promotion behavior:

H3: *Semantic differences in review promotion ("helpful" vs. "like") influence how users promote reviews.*

In a controlled experiment on CrowdFlower[2], participants were asked to rate reviews. In one treatment participants had

[2]crowdflower.com

Reasons	H	¬H	L	¬L
It provided an objective point of view.	14.0	23.6	12.5	25.3
It helped me decide whether to buy the book.	42.3	11.0	38.2	11.5
It helped me learn more about the book.	38.0	51.1	38.7	43.5
It was enjoyable to read.	3.6	11.4	6.6	12.9
I agreed with the reviewers point of view.	2.0	2.0	3.7	5.2
Other.	0.1	0.9	0.3	1.6

Table 3: Percentage of reasons given for applying each label to reviews. Each column sums to 100%. **H** = "helpful"; ¬**H**= "not helpful"; **L** = "liked"; ¬**L** = "not liked";

to "like" or "dislike" reviews, while in the other participants chose between "helpful" or "not helpful". The same reviews were used in both treatments. Participants were asked to select a reason for their judgement (see Table 3). All reviews were marked by six different workers: three for each treatment. Workers had to select one of the options for each review, to get a large enough number of marked reviews to be analyzed. Workers could evaluate at most 100 reviews each. We selected equal numbers of reviews having been promoted (or not) from both platforms, taking 476 in total. We randomly inserted 60 *test reviews* per job, which were negative by construction consisting only of numbers (no text), random words, or Latin filler text. Workers rating a test review positively were disqualified and any prior judgements discarded.

Concepts comparison Workers chose "like" and "helpful" in similar proportions (70% and 76% respectively), and gave reasons for both in similar proportions (Table 3). The most common reasons were that it helped the person decide whether or not to buy the book, and that it helped them learn about the book, suggesting informativeness is an important factor to both concepts.

To measure the similarity in review promotion, we count the number of "like" and "helpful" votes for each review, then measure agreement using Krippendorff's alpha. Doing so reveals a modest agreement of 0.22, showing the concepts are correlated but not strongly. We test whether these differences are significant using the Stuart-Maxwell test for marginal homogeneity. The test shows that the concepts are, statistically, highly distinct ($\chi^2 = 55.7$, $\mathrm{d.o.f.} = 3$, $p < 0.001$), showing that different semantics causes users to consider different factors when deciding whether to promote a book review.

Discussion

We have made the case that the Amazon and Goodreads platforms yield different reviewing behaviors. We have shown substantive differences in all three core areas of review engagement: review content, star-ratings, and review promotion. This favors the conclusion that platforms influence reviewing practices. But what produces these differences?

Our crowd sourcing experiment highlights the impact that the design of a review feedback feature can have on which reviews are promoted. The observed differences in review writing may well be attributable to these kinds of effects.

It remains possible that differences in review content arise from platform-driven *selection*. The platforms may preferentially attract users of different age or level of education. Alternatively users might be influenced by the platforms' design and existing content. It is likely a combination of both. Efforts to disentangle selection and influence are stymied by the absence of users' demographic information. But, even with such information, the feedback between social influence and self-selection would be difficult to unwind.

It is nevertheless clear that these platforms yield different styles of reviews despite shared subject material. These findings have implications for users seeking in depth reviews, and for the design of interfaces that elicit content from users.

References

Diakopoulos, N., and Naaman, M. 2011. Towards quality discourse in online news comments. In *CSCW*.

Dimitrov, S.; Zamal, F.; Piper, A.; and Ruths, D. 2015. Goodreads versus amazon: The effect of decoupling book reviewing and book selling. In *ICWSM*.

Forman, C.; Ghose, A.; and Wiesenfeld, B. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19(3):291–313.

Gilbert, E., and Karahalios, K. 2010. Understanding deja reviewers. In *CSCW*.

Liu, J.; Cao, Y.; Lin, C.-Y.; Huang, Y.; and Zhou, M. 2007. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*.

Lui, M., and Baldwin, T. 2012. langid. py: An off-the-shelf language identification tool. In *ACL system demonstrations*.

Michael, L., and Otterbacher, J. 2014. Write like i write: Herding in the language of online reviews. In *ICWSM*.

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

Otterbacher, J. 2009. 'helpfulness' in online communities: a measure of message quality. In *CHI*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71:2001.

Sukumaran, A.; Vezich, S.; McHugh, M.; and Nass, C. 2011. Normative influences on thoughtful online participation. In *CHI*.