Understanding Online Reviews: Funny, Cool or Useful?

Saeideh Bakhshi
Yahoo Labs
San Fransisco, CA
sbakhshi@yahoo-inc.com

Partha Kanuparthy
Yahoo Labs
Sunnyvale, CA
parthak@yahoo-inc.com

David A. Shamma
Yahoo Labs
San Fransisco, CA
aymans@acm.org

ABSTRACT
Increasingly online reviews are relied upon to make choices about the purchases and services we use daily. Businesses, on the other hand, depend on online review sites to find new customers and understand people’s perception of them. In order for an online review community to be effective to both users and businesses, it is important to understand what constitutes a high quality review as perceived by people, and how to maximize quality of reviews in the community. In this paper, we study Yelp to answer these questions. We analyze about 230,000 reviews and member interaction (“votes”) with these reviews. We find that active and regular members are the highest contributors to good quality reviews and longer reviews have higher chances of being popular in the community. We find that reviews voted “useful” tend to be the early ones reviews for a specific business. Our findings have implications on enabling high quality member contributions and community effectiveness. We discuss the implications to design of social systems with diverse feedback signals.

Author Keywords
Yelp; Online Reviews; Social Signals; Votes; Social Feedback; Zero Inflated Negative Binomial Regression; Funny, Cool, Useful Votes

ACM Classification Keywords
H.4 Information Systems Applications: Miscellaneous; D.2.8 Software Engineering: Metrics—complexity measures, performance measures

General Terms
Human Factors; Measurement.

INTRODUCTION
Online review sites such as Yelp are important and widely-used resources that enable members to share their experiences with products, services and activities in the form of reviews and ratings at scale—which may otherwise be difficult to ascertain before receipt of the service or product. Over 138 million unique users visited Yelp in the mid 2014 and had over 61 million reviews1. Research has shown that consumers perceive online review sites as unbiased sources of information when compared with business websites [10].

Online review sites are only as good as the content their users provide. It is critical for an online community to provide mechanisms that encourage contributions from members, and at the same time, and perhaps even more important, to have high quality contributions that engage users. Unfortunately, the abundance of user-generated content comes at a price: for every interesting opinion or helpful review, there are content and opinions that are unhelpful, subjective or misleading. Sifting through large quantities of reviews to identify high quality and useful information is a tedious and error-prone process. Online review sites typically use crowd-sourced methods to rank reviews. An example of such a social evaluation mechanism is the Yelp review votes, that enable members to flag a review as funny, cool and/or useful. Understanding social evaluation of reviews can aid review sites design mechanisms to improve the quality of contributions.

Yelp is an online community where people review and rate businesses. Users of Yelp can search for businesses using keywords. Yelp allows users to interact with each other by voting others’ reviews, and by following activity of their friends. In this work, we ask three research questions: (i) What are the factors behind engaging reviews? (ii) What is the nature of reviews voted funny, cool or useful (social feedback)? and (iii) How do these votes relate to user ratings of establishing quality?

We use a large dataset from Yelp from the Greater Phoenix metropolitan area in the USA. We study 43,873 reviewers and their 229,907 reviews of 11,537 businesses, spanning 2005 to 2013. While our data and analysis is specific to Yelp, the implications and research questions we answer are general and may apply to online review communities in general.

We find that members who are active for longer periods of time tend to be more significant contributors of quality reviews. We find that reviews that members find funny tend to be negative in tone. We also find that there is a direct relationship between social evaluation and individual evaluation: funny reviews tend to have low contributor ratings, while cool reviews tend to have high ratings. Our results also suggest that readers tend to like long and objective reviews.

Our study highlights insights on online review communities, and brings implications for design of recommendation communities to promote high quality contributions by members.

1http://yelp.com/about
We show mechanisms that can encourage longer reviews, provide incentives for new members to contribute and incentives for members to review new businesses, and significantly improve quality of contributions and effectiveness of the community. Further, our findings on Yelp's social feedback system inspires design of distinct signals in other social platforms.

RELATED WORK

General aspects of online user engagement have been discussed in detail in prior work [18, 26]. One common way to study user engagement is through peer evaluation. For a user, feedback from fellow members could lead to future participatory behavior. Theories of reciprocity [8, 13], reinforcement [21], and the need to belong [4] suggest that feedback from other users should predict long term participation on the part of the users. For example, users of the online news community Slashdot whose first comments received positive numeric ratings returned significantly faster to the site to post a second comment, and when their first comment received a reply, they also tended to return more quickly [17]. Controlled experiments also show that social approval in the form of messaging increases a user's number of contributions [7].

To encourage both creators of content and the readers to engage with the site, many online communities provide users with a feedback system. Facebook likes, Twitter’s favorites or retweets, Amazon’s helpful votes on reviews and Yelp’s useful, cool and funny votes are all examples of such systems. Prior research suggests that perceived attributes of the review text, reviewer and social context may all shape consumer response to reviews [1, 19]. It has been shown that even exogenous factors such as weather and demographics of users might impact the ratings and reviews [2].

There is a body of work on analyzing product reviews and postings in forums. Lu et al. use a latent topic approach to extract rated quality aspects from comments in eBay [19]. Another work looked at the temporal development of product ratings and their helpfulness and dependencies on factors such as the number of reviews or effort required (writing review vs. just assigning a rating) [25]. A 2008 work looked at the helpfulness of answers on the Yahoo Answers site and the influence of variables such as required type of answer and the topic domain of the question [14]. A study on Amazon reviews looked at the helpfulness scores and found that the helpfulness scores are not only dependent on the content of the review but also on other reviews posted for the product [9]. More recently, we ran a study on Yelp, to indentify whether the social signals of a review are indicative of review’s rating and sentiment [3]. While there have been studies on understanding and predicting helpful votes, we do not know much about what factors shape the other social signals such as cool and funny votes on Yelp.

DATA

We use a publicly released dataset from Yelp to answer our research questions\(^1\). Yelp is a large online review community that is also a member-maintained business and service directory. The Yelp dataset consists of a sample of Yelp data from the Greater Phoenix Arizona metropolitan area. It includes 11,537 businesses, 43,873 reviewers and 229,907 reviews. The data spans 2005 to 2013.

Businesses listed on Yelp have three prominent review attributes that a user sees: an average “stars” rating (scale of one to five), the number of reviews and the reviews themselves (along with ratings). We process the review timestamps for a business to measure the active period, defined as the duration between the first and last reviews written on Yelp for that business. We use active period to account for time when a business was not operational (e.g., if it was established after our dataset started, or ended before the dataset finished).

As in any online review community, Yelp members contribute reviews for businesses. A review consists of a stars rating and a review text. In order to understand the effect of review text, we process the text for each review to quantify the subjectivity (between 0 and 1) and polarity (between -1 and 1) of the text using the Pattern Toolkit [22]. We have used the number of words in the review as a measure of the length of the review. Yelp users can vote a review as one or more of cool, funny and useful. We use these vote counts, along with total number of votes, as a measure of review social feedback.

We quantify reviewer activity on the site by looking at the number of reviews the reviewer writes, and the votes (cool, funny and useful) those reviews get from other Yelp users. We also consider the average stars the reviewer uses in her reviews. We process a user profile to compute two metrics: (i) activity duration, defined as the duration between the first and last review written on Yelp and (ii) activity rate, defined as the number of reviews written divided by the activity duration. For example an activity rate of 0.2 implies that on average, the reviewer wrote one review every five days during the active period.

We expect that demographics in the location of a business can have an impact on reviews and ratings of a business [2]. All businesses in our data are from the same metropolitan area; however, they are located in different neighborhoods (qualified by latitude-longitude), each with its own demographics. We analyze the effect of user diversity by including demographic factors in the neighborhood of the business. For each business, we collect two dimensions of demographics at the business location (latitude-longitude or neighborhood). First, we collect the median income for residents in the location. Second, we collect the education level for that location, defined as the fraction of residents who have a bachelors degree.

---

\(^1\)https://www.yelp.com/dataset_challenge
or higher. This was collected from the US National Broadband Map for demographics data.  

FACTORS SHAPING SOCIAL FEEDBACK  

There are multiple ways to study the social feedback system on Yelp. These feedback signals provide three different utilities for users: funny, useful, and cool signals. We utilize statistical methods to understand the relationship between review, user and business attributes with these signals; this allows us to understand the relationship between votes given by users to reviews and the reviews themselves. If votes are distributed randomly by people on site, then there would be no consistent relationships. If we find systematic relationships, this suggests votes are used and can be used to evaluate reviews.

We use regression models that take the number of votes (total number of votes as well as cool, funny and useful votes) as their dependent variables; and consider business, review and reviewer features as independent variables. The number of votes is an overdispersed, distributed count variable (See table 1) with a probability function for a zero-inflated negative binomial regression model [6]. We study two types of independent variables: the variables describing the attributes of the review and the variables describing the attributes of the reviewer writing the review. We also consider variables related to the business being reviewed to control for the factors such as business active days and demographics around the business. Table 2 shows the coefficients of our zero-inflated negative binomial regression model for the count model. We consider the count model and effect on reviews with non-zero votes. The regression coefficients allow us to understand the effect of an independent variable on the number of votes (note that to be able to compare coefficients, we z-score all numerical variables before regression).

We use the Chi-square Test to find the statistical significance of our regression models, by computing the reduction in deviance from a null model. For our model for the total number of votes, we found the reduction in deviance $\chi^2$ of $(447.1 \times 10^3 - 390.9 \times 10^3)$, or 13%, for 27 degrees of freedom. The test rejected the null hypothesis of a null model ($p < 10^{-15}$); hence, the regression model is well-suited to characterize the effects of the independent variables. We test coefficients of all independent variables for the null hypothesis of a zero-valued coefficient (two-sided) and found that the test rejects the null hypothesis ($p < 10^{-3}$) in all cases.

Active reviewers are more likely to write high quality reviews.

The regression model shows that the activity level of a Yelp user has a significant relationship with the review quality (Table 2). Activity can be quantified as either the number of reviews the reviewer has written (coefficient $\beta = 0.22, p < 10^{-15}$) or the duration the user has been reviewing on Yelp ($\beta = 0.41, p < 10^{-15}$). We can explain this relationship through the knowledge and experience an active user gains from the Yelp community: active users have the advantage of gaining knowledge about the community and its interests. Such knowledge may reflect on the quality of reviews experienced members write. Prior studies have shown that one of the reasons people join online communities is to access information and gain knowledge [16, 24].

We see that the average number of stars a reviewer tends to give in her reviews has a small positive coefficient for the number of votes ($\beta = 0.05, p < 10^{-15}$). In other words,
a user who rates businesses higher might have a slight advantage in receiving votes compared to a reviewer who rates lower. Moreover, activity level of a user is also related to the number of votes the user’s reviews receive. One way to explain this relationship is through theories of social identity [11, 15, 23], according to which people form a social identity of values, attitudes and behavioral intentions from the perceived membership in social groups. On the other hand, one could argue that the feedback system might be the encouraging factor to more participation. We know from previous work that theories of reciprocity [8, 13], reinforcement [21], and the need to belong [4] suggest that feedback from other users predicts long term participation on the part of the users.

Longer and objective reviews are the main identifiers of high quality reviews.

The content of a review and how much information it provides can be one of the reasons why reviews get more votes. Indeed, our model shows that the length (number of words) of a review plays a significant role in the number of votes that the review gets ($\beta = 0.3$, $p < 10^{-15}$). This may be explained by the hypothesis that longer reviews are likely to contain more information about the business, and such reviews are likely written by more dedicated reviewers. In the next section, we show that longer reviews get higher number of useful, cool and funny votes. Our observation on longer reviews also shows that Yelp users tend to prefer to read longer reviews.

Interestingly, certain features of the review text have a negative effect on the number of reviews. Our sentiment analysis shows that sentiment polarity of the review text has a strong negative relationship with the number of votes ($\beta = -0.12$, $p < 10^{-15}$). We also find that subjectivity of the review text has a small but negative relationship with the number of votes ($\beta = -0.01$, $p < 10^{-5}$). These observations show that users perceive objective and less polarized reviews as higher quality than subjective and more polar ones. The coefficient for user rating is positive but small compared to other review features ($\beta = 0.02$, $p < 10^{-3}$).

COOL, FUNNY, USEFUL VOTES

In the previous section, we modeled the total number of votes a review gets and studied the factors shaping review quality. In this section, we try to understand how users perceive (vote) a review. Specifically, what are some differences among funny, cool and useful reviews?

We construct zero-inflated negative binomial regression models for funny, cool and useful votes of a review as a function of reviewer, review and business-related features. The independent variables in our models are similar to those used in the previous section for modeling total number of votes. For all models, we test statistical significance using a Chi-squared Test on reduction of deviance. We find that all tests reject the null hypothesis of a null model ($p < 10^{-15}$). Hence, the models are well suited to characterize the effects of the independent variables. We test coefficients of independent variables for the null hypothesis of a zero-valued coefficient (two-sided) and find that the test rejects the null hypothesis ($p < 10^{-3}$) for all variables.

Funny, cool, and useful votes are correlated with the reviewer’s activity and average stars.

We find that the number of reviews a user writes has a strong relationship with all three types of votes that the user’s review gets ($\beta_{\text{cool}} = 0.53$, $\beta_{\text{funny}} = 0.54$, $\beta_{\text{useful}} = 0.33$). In addition, we see that a regular Yelp reviewer has higher chances of writing funny or cool-voted reviews; however, the chances of getting useful votes are lower. This suggests that the reviewer’s experience has a smaller impact on writing useful reviews compared to writing funny or cool reviews (as perceived by other users). Another explanation could be that while useful votes are given to informative reviews for the wider set of Yelp readers, cool and funny votes are more community oriented as hence are more likely to be given to Elite users with higher number of reviews.

The reviewer activity duration has a similar relationship with different types of votes as the number of reviews ($\beta_{\text{cool}} = 0.25$, $\beta_{\text{funny}} = 0.23$, $\beta_{\text{useful}} = 0.18$). This can probably be explained using the hypothesis that older users in the community understand the community and the notions of cool, funny and useful better than new users.

We also find that the average stars a reviewer gives in her ratings is positively correlated with the number of cool, funny and useful votes that her review gets ($\beta_{\text{cool}} = 0.16$, $\beta_{\text{funny}} = 0.04$, $\beta_{\text{useful}} = 0.04$). Further, we find that this correlation is stronger with cool votes compared to funny and useful votes. This suggests that reviewers who rate businesses higher have higher chances of writing reviews that are perceived cool. This could be explained by more positive expectations from cool reviews than from funny ones, given our findings that funny reviews tend to a more negative tone.

Early reviews are usually the most useful ones.

We see that highly reviewed businesses are more likely to receive funny, cool and useful votes ($\beta_{\text{funny}} = 0.08$, $\beta_{\text{useful}} = 0.07$, $\beta_{\text{cool}} = 0.09$). This effect is intuitive, we expect that the restaurants with higher number of reviews are more popular, and so more visitors check out their profiles and reviews on Yelp. As a result, the average number of votes on those businesses are expected to be higher than the less reviewed businesses. We also find a small but positive effect of business review stars ($\beta_{\text{cool}} = 0.02$, $\beta_{\text{funny}} < 10^{-2}$, $\beta_{\text{useful}} = 0.03$).

The active duration for a business has a positive effect on funny and cool votes ($\beta_{\text{funny}} = 0.03$, $\beta_{\text{cool}} = 0.02$) but a negative effect on useful votes ($\beta_{\text{useful}} = -0.04$). This observation suggests that the businesses that have been on Yelp for a long time are not likely to receive useful votes, but could receive funny and cool votes. This is a possible indicator of maturity of an online community and we find it an interesting observation for a successful community such as Yelp. In a recent work, Gilbert asked a related question [12]: why do reviewers write reviews when there are already enough useful reviews? He found that an overwhelming number of reviewers who write deja reviews look for individual status in the online community.
**Funny reviews are more negative in tone.**

We find that the length (number of words) of a review has a strong correlation with the number of funny, cool and useful votes it gets ($\beta_{\text{cool}} = 0.27$, $\beta_{\text{funny}} = 0.29$, $\beta_{\text{useful}} = 0.27$). This can be explained by considering that longer reviews are likely to have higher information content about the business.

We find a disparity in the coefficients of review stars (ratings) on the different types of votes the review gets. Review stars has a small negative relationship with the funny and useful votes ($\beta_{\text{funny}} = -0.04$, $\beta_{\text{useful}} = -0.01$). However, the relationship of stars is positive and much higher in magnitude with cool votes ($\beta_{\text{cool}} = 0.18$). In other words, higher rated reviews are likely to get cool votes, but are unlikely to get useful or funny votes.

We find that review text is related to all three types of votes. We see that both polarity ($\beta_{\text{cool}} = 0.16$, $\beta_{\text{funny}} = 0.04$, $\beta_{\text{useful}} = 0.04$) and subjectivity ($\beta_{\text{cool}} = 0.16$, $\beta_{\text{funny}} = 0.04$, $\beta_{\text{useful}} = 0.04$) of a review have negative impact on funny, cool and useful votes that the review gets. In addition, polarity has a strong negative relationship with funny votes, showing that polarized and subjective reviews are not likely to be perceived as funny by users. This can be explained in two ways. First, it might be that the general audience on the review sites such as Yelp enjoy reading reviews with “sarcastic” tone or those reviews that criticize the business or service with humor. The second hypothesis could be that users find the lower rated reviews and those with negative tone more funny or humorous. Either way, the observation that higher funny votes are related to the lower rated reviews illustrates an artifact of the what the Yelp user considers funny. Additionally, one could argue that funny votes are used as a negative signal to express dislike on the review—since the other voting signals on Yelp have positive connotations, users may have utilized the funny signal to express negative emotions. Unlike the perceptions of funny and useful on the reviews, we also found that reviews that are perceived as cool are more likely to be highly rated. This suggests that the cool perception is usually tied with the higher rated reviews. The lower rated reviews are less likely to be perceived as cool.

**IMPLICATIONS AND FUTURE WORK**

On Yelp, funny, cool, and useful votes are not random or whimsical signals to attract click engagement, but are actually good measures of quality—quality that is expressed in different ways. Online review communities can rely on member contributions to index and serve recommendations to their users. It is critical to maximize the articulation and the quality of these contributions.

Most social networking sites use like, favorite or upvote signals to capture users interest in a piece of content. Product sites like Amazon allow users to vote reviews as “helpful”. Similar to social networking sites, online review communities have adopted various signals to improve user engagement and crowdsourcing quality control. However, these signals can carry different meanings within themselves. For example, someone can like a photo on Facebook because she likes the person in the photo, or because the photo captures a beautiful scene. Yelp, on the other hand distinguishes the votes by dividing them into three different signals. Each of these signals carry different meanings beyond their specific labels.

Our findings suggest that there are deeper meanings and interaction forms than the generic measure of votes. The Yelp community, for example, judges and communicates meaning through its own interpretation of the signal labels, which is similar to findings in the multimedia community [20]. Future work can elaborate on the use of these labels and their corresponding social perceptions in other communities. For example, what do likes on Facebook mean in different contexts and by different people? Is it some form of social confirmation? Do users support the content of the post? How does it relate to Yelp’s cool and funny signals?

Deeper understanding of forms of user feedback can enable designing better recommendation and social networks. For example, mechanisms for users to search for useful reviews or cool reviews or providing them with ranking options based on the number of votes. On the other hand, in social networks with more broad content types, new feedback signals might enable users to interact with the content in a more meaningful way.

Based on our findings, sections of active members can be leveraged to improve contributions by other members. For example, old timers who are active in the community tend to be the people who write the most reviews and get the most votes for their reviews. This is an important finding, since such users are popular in the community, and social mechanisms can be designed to encourage interaction with newer members and as a consequence, quality contributions by newer members.

We showed that longer reviews get more user votes. Community mechanisms to encourage review length or reward reviewers who write longer reviews may help improve review quality. At the same time, such mechanisms should be carefully designed so that they do not discourage reviewers who have a tendency to write short (but useful) reviews from not writing reviews.

New businesses added to an online review community can benefit from initial opinions, especially from old timers, to help gather “review momentum”. We found that new businesses have the highest chances of receiving useful reviews in the initial period of their appearance on Yelp. To improve the quality of reviews for such businesses, the community can include incentives for old timers and experienced reviewers to write reviews for new businesses. Such users are influential and can set a trend for better quality reviews for new businesses in the community.

Finally, design of social feedback signals such as Yelp funny, useful and cool votes can be used to encourage conversations around a piece of content. As Brown [5] suggests in his work, review sites are beyond search and recommendation, and can be starting points to conversations around local businesses. By voting a review as useful, the reader conveying that she liked the review and likes to see more of such reviews. While this can be used in personalizing search and recommendation, it can also start a new conversation between the voter and the
CONCLUSION
Online review communities such as Yelp enable people to find right businesses and services, and enable businesses to find new customers and improve profit margins. Hence, it is critical for such communities to maximize high quality member contributions. In this paper, we studied member interaction, reviews and businesses in the Yelp online recommendation community. We found that active and older members tend to contribute significant content. We found several factors that contribute to recommendations that are perceived as useful, cool or funny. For example, we found that reviewers who give higher ratings in their reviews tend to be perceived as writing cool reviews. We found that longer reviews are perceived as useful, cool and funny. We saw that reviews that are negative in tone are more likely to be seen by users as funny. We showed that early businesses listed on Yelp tend to get reviews with higher ratings. We found that businesses listed on Yelp tend to get reviews that are perceived as useful. Our findings have implications on online review community design for improving quality of member contributions.

REFERENCES
