The Social Aspect of Voting for Useful Reviews

Asher Levi[1] and Osnat Mokryn[2]

[1] Department of Information Systems, University of Haifa, Israel
[2] School of Computer Science, Tel Aviv Yaffo College, Israel

Abstract. Word-of-mouth is being replaced by online reviews on products and services. To identify the most useful reviews, many web sites enable readers to vote on which reviews they find useful. In this work we use three hypotheses to predict which reviews will be voted useful. The first is that useful reviews induce feelings. The second is that useful reviews are both informative and expressive, thus contain less adjectives while being longer. The third hypothesis is that the reviewer's history can be used as a predictor. We devise impact metrics similar to the scientific metrics for assessing the impact of a scholar, namely *h-index,* i_5*-index*. We analyze the performance of our hypotheses over three datasets collected from Yelp and Amazon. Our surprising and robust results show that the only good predictor to the usefulness of a review is the reviewer's impact metrics score. We further devise a regression model that predicts the usefulness rating of each review. To further understand these results we characterize reviewers with high impact metrics scores and show that they write reviews frequently, and that their impact scores increase with time, on average. We suggest the term *local celebs* for these reviewers, and analyze the conditions for becoming local celebs on sites.

1 Introduction

In today's Internet, reviews for products and services have become an online form of word-of-mouth. Reviewers contribute their time and energy to express their opinions and share their experience. Shoppers, as well as online service consumers, read these reviews as a first step in their decision-making process. With the growth in popularity of online reviews in popular commerce sites like Amazon, numerous sites have been created for the sole reason of becoming review portals: Yelp (businesses and services reviews), TripAdvisor (hotel reviews), and more.

Browsing through the plethora of reviews, readers find it hard to navigate and extract the useful information. Hence, a system that enables one to determine which reviews are more useful, or helpful, is needed. Sites like Amazon, Yelp and IMDb thus rely on a form of crowdvoting, in which the readers specify which reviews they find useful, or helpful. Specifically, Amazon customers and IMDb users vote whether a review is helpful or not; In Yelp, the users up-vote to indicate they find a review useful.

The purpose of this paper is to characterize which reviews are voted useful [3]. The benefits of automating this selection process are numerous. It enables the promotion

[3] The term useful refers also to helpful reviews. Throughout this paper, these terms are used interchangeably.

of potential useful reviews thus presenting users with current useful reviews upon their arrival, to improve the shopping / browsing experience; Useful reviews can be detected early, enabling sites to keep an up-to-date and current review system, without compromising quality; Useful reviews for low-traffic or new items can be detected, while today they might not gather enough votes to be noticed. Research in this area in the last years focused on classifying which reviews are voted useful or ranking useful reviews [1,2,3,4], taking into account textual characteristics, review sentiment, or product and user features. We continue to suggest three novel hypotheses for why reviews are voted useful.

Feelings have shown to influence decisions [5,6]. Physiologists have long demonstrated that people can induce feelings in others [7]. For example, imagine you enter a room in which a friend converses happily, laughing. Experience shows that you are likely to find yourself smiling at ease. Does a similar effect exist in text? Can reviews induce feelings and influence a reader's decision and therefore be perceived as more useful? We then come to our first hypothesis: (1) Useful reviews are more likely to express feelings.

Useful reviews are perceived as carrying important information to readers. Informative text contains more nouns, while descriptive text contains more adjectives. Additionally, longer text carries more information and might therefore be perceived as more useful. Our second hypothesis is hence: (2) Useful reviews are longer and more informative yet expressive (e.g., high ratio of nouns per adjectives accompanied with punctuation marks).

Profiling the reviewers reveals that some people tend to have a large number of useful reviews. As writing a review might be seen as a skill, the question is can skilled reviewers be identified, and are they more likely to keep producing useful reviews. To that end we use the average useful count a reviewer received. However, as the different sites employ different voting and counting systems, we proceed to create a set of unified metrics for capturing a reviewers "useful" counts for previous reviews. We thus devise a common metric for profiling a reviewer, borrowing from the scientific metrics for measuring a scholar's impact: the famous *h-index* and i_x*-index*. Our hypothesis is then: (3) A reviewer's impact score is a good predictor for the perceived usefulness of her future reviews.

We evaluate our hypotheses on three different datasets. Two datasets are from Yelp and one is from Amazon. All three data sets are cross domain, and are detailed in Section 3. Our results (detailed in Section 4) are conclusive, showing that the reviewer impact score is an excellent predictor, outperforming all others. Specifically, a SVM classifier with these users features gives an outstanding 97.8 accuracy (with AUC of 0.97) for the Amazon datasets, and almost as good results for the Yelp datasets. Moreover, our exact vote count prediction R-square values range between 0.561 and 0.871 across all evaluated datasets, with an RMSE of less then 1.7% of the predicted values. We further checked the statistical significance of each feature across the three datasets.

In Section 5 we explore the origin of these findings. We find a tendency for reviewers with high *h-index* to author many reviews. We further found that new reviewers are less likely to receive a high number of useful votes. Specifically, later reviews of the same author would on average be perceived as more useful and would be up-voted more

as useful. We suggest the following explanation: Research on online engagement has demonstrated the importance of trust. An up-vote is an expression of trust, as the voter would only indicate as useful reviews she trusts their content to be truthful and accurate. It has been demonstrated that familiarity is a fundamental perquisite for trust, especially online [8,9]. Reviewers investing time and energy in writing reviews become known in online communities, in what we term *local celebs*, and their reviews are therefore perceived as useful.

2 Modeling and Hypotheses

A common review system is typically comprised of three different types of entities: a set $I = i_1, ..., i_N$ of N items (businesses, products or services); a set $U = u_1, ..., u_M$ of M users (or reviewers) that use the system and write reviews; and finally, a set $R = r_1, ..., r_T$ of T reviews that were written by the users, expressing their opinion on these items. Each review r written by reviewer r on item i contains the following information: users rating, review text, and the number of useful votes v. Formally, given a set of reviews R_i for a particular item i (business, product or service), we try to: Classify whether a review is useful or not; Predict the count of useful votes v for each review. The formal definition for a useful review is given in Section 3.

Our experience with reviews [10] has demonstrated that the text of the reviews contains additional information beyond the rating given by the reviewer. Consequently, it seems natural to assume that reviews perceived as useful would contain information found important by users. A survey of useful reviews led to the following observations:

1. Reviews expressing emotions are more likely to induce feelings, and hence be perceived as useful.
2. Useful reviews tend to be more informative (for example, have a higher nouns/adj. relation), and often contain expressive text expressions.
3. Users who historically have written useful reviews write more useful reviews.

An explanation of the text emotions and sentiment hypotheses can be found in the extended version [11], we concentrate here on modeling reviewer characteristics. Our observations have led to the conclusion that it is likely that the reviewer's information may also be a predictor for whether a review will be voted useful and how many votes it will acquire. It is reasonable to assume that some people are better than others in writing reviews, and hence their reviews will be perceived as more useful. Indeed, many sites display the reviewer information in a prominent way. In Yelp, for example, "Elite members" (top reviewer) are elected each year, and earn a badge that is emphasized online. In Amazon the "Top-1000" reviewers have a special tag displayed next to their name. The challenge ahead is to examine whether the reviewer's history can be used to predict the useful votes of her future reviews. To that end, a metric for describing the history of a reviewer needs to be formalized.

Reviewer Impact Metric A reviewer profile usually consists of the following: number of reviews, average useful votes, sum of useful votes, star rating, and user average star rating.

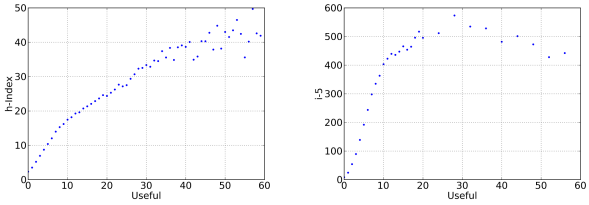
To capture the *perceived impact* of a reviewer we take a detour to known methodologies

to assess the impact of scientists and scholars. The *h-index* is a widely known and heavily used metric in science and academia. From Wiki: *The index is based on the distribution of citations received by a given researcher's publications, ... and reflects both the number of publications and the number of citations per publication.* Google Scholar further calculates a second metric, i_{10}*-index*, which describes the number of papers with 10 or more citations the scholar has published. Based on the above metrics, we define two variables to capture a reviewer's history: *h-index*, i_5*-index*.

Reviewer *h-index* : Reviewer has index h if h of her N_p reviews have at least h useful votes each, and the other $(N_p - h)$ reviews have no more than h useful votes each. Hence, an *h-index* of 10 means the reviewer has at least 10 reviews which were voted useful by at least 10 other people, and the rest of her reviews have less than 10 useful votes each.

Reviewer i_5*-index* : Reviewer has index i if she has i reviews with at least 5 useful votes each.

Fig. 1 shows the average number of useful votes users with the same *h-index* and i_5 *index* receive. Clearly, the better is a reviewer's impact, the more useful votes her reviews will receive on average. The complete list of features can be found here [].



(a) *h-index* vs. useful vote (b) i_5 *index* vs. useful vote

Fig. 1: Average Useful Votes vs. Average Reviewer Metrics

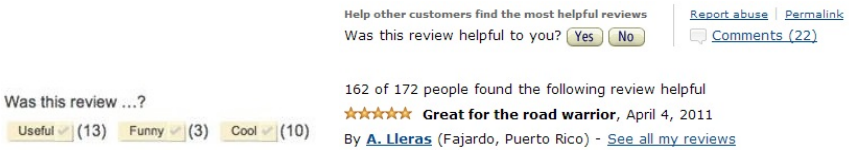
3 Data

To evaluate our hypotheses we use three different datasets, as depicted in Table 1. Two data sets contain reviews about businesses from YelpThe first Yelp data set, referred to as *Yelp Bay Area*, was collected from Yelp in the Bay area, California, during May and June 2011. The second data set, referred to as *Yelp challenge*, was introduced in early 2013 by Yelp. It includes businesses' reviews from the greater Phoenix metropolitan area. The third data set, referred to as *Amazon*, is a set of product reviews obtained from Amazon during April 2010. Yelp users vote whether a review is useful, while Amazon users vote whether the review was helpful or not. (see Figures 2a, 2b). Using these votes, the community can filter relevant opinions more efficiently.

First, we need to convert the continuous variable "Useful" in Yelp, or "Helpful" in Amazon into a binary one. For the Amazon dataset, *Helpful* is a vote h out of v, where h is number of helpful votes, and v is the total number of votes. A review is marked as useful when h/v is greater then the threshold γ, and h is greater then threshold ζ. [1]

Data Set	Users	Items	Reviews	Useful	%Useful
Amazon	21087	2103	23868	12877	53.9%
Yelp Bay Area	95296	66803	488805	63074	12.9%
Yelp challenge	43873	11537	229907	14422	6.2%

Table 1: Summary of statistics for each of the data-set



(a) Yelp review useful voting (b) Amazon review helpful voting

Fig. 2: Useful and Helpful Votes in Sites

showed that the best threshold γ is 0.6. We chose γ as 0.6 and ζ as 5. In other words, if a review has more then 5 votes indicating that the review is helpful, and more than 60% of the votes indicate that the review is helpful, then we classify the review as "useful". Otherwise, we classify the review as "non-useful".

For the Yelp datasets *useful* is a vote h, where h is number of useful votes. Review is marked as useful when h is greater then the threshold ζ. The same ζ was chosen as for the Amazon dataset. In other words, if the review has more then 5 votes indicating that the review is useful, we classify the review as "useful". Otherwise, we classify the review as "non-useful".

4 Experiments and Results

We have conducted two large scale experiments to explore the validity of our hypotheses, as detailed in Section 2. A classification, and a useful score prediction. In both experiments, a review r is represented as an f-dimensional real vector x over a feature space F constructed from the information in r. Our methodology is based on a 6-fold cross validation procedure. To classify whether a review is useful or not, we use a supervised learning paradigm, based on a training set containing labeled data. The test set contains reviews represented as multi-dimensional feature vectors. We employ Support Vector Machines (SVM), using as kernel the radial basis function (RBF). The evaluation metrics are the classification accuracy and the area under the ROC curve (AUC)

We performed different series of binary classification experiments for evaluating our different hypotheses separately and combined, as detailed here. A detailed features list can be found in our extended version [11]. The first experiment, referred to as *All*, uses all the features; We then proceed to validate each one of our hypotheses as detailed. For evaluating our hypothesis on induction of emotions, referred to as *Emotions*, we used all the emotions features. For the text hypothesis, referred to as *Text*, we use the text features; We evaluate the combination of the two hypotheses above in an experiment referred to as *Emotions-Text*, using the combined set of features from *Emotions, Text*. Our third hypothesis, which takes into account the reviewer's impact, is referred to as

Experiment	Yelp Bay Area		Amazon		Yelp Challenge	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
All	92.5%	0.79	97.7%	0.97	95.3%	0.68
Emotions	87.1%	0.5	59.5%	0.58	93%	0.5
Text	87.1%	0.5	60.1%	0.59	93.7%	0.5
Emotions-Text	87.1%	0.5	61.4%	0.61	93.7%	0.5
Reviewer	**92.3%**	**0.8**	**97.8%**	**0.97**	**95.2%**	**0.69**

Table 2: Accuracy and AUC for the Usefulness Classifiers

Dataset	R-squared	RMSE	Useful Range	RMSE(%)
Yelp Bay Area	0.607	3.02	$0 - 180$	1.67%
Amazon	0.871	7.56	$0 - 635$	1.19%
Yelp Challenge	0.561	1.47	$0 - 120$	1.22%

Table 3: R-squared and RMSE for the Prediction Model

Reviewer. Table 2 shows the results for the experiments over all the datasets described in the previous section.

The first observation noted is that the experiments *Emotions*, *Text*, and *Emotions-Text* show low performance. The prediction accuracy for those experiments is high, while the AUC is low; Further analysis reveals that the precision of the classifier is high but the recall is low; This, with the fact that there is a low rate of useful reviews in the Yelp datasets (12.9% and 6.2%), may indicate that the prediction only succeeds in predicting "non-useful" reviews but fails in predicting the useful ones. The results for these experiments on the Amazon dataset support this observation: As the share of useful reviews in the Amazon dataset is much higher (53.9%), we see that indeed both the accuracy and the AUC are low. Hence we conclude that the results don't support our two first hypotheses.

The Reviewer Characteristics features set achieved the best performance across all datasets, if we take into account both the accuracy and AUC. Specifically, it outperformed all other feature combinations on the Amazon dataset, where there is a larger percentage of useful reviews, and performed similar to All Features experiment on both Yelp datasets. Hence, Reviewer Characteristics show to be a high quality and excellent predictor for the usefulness of a review, supporting our third hypothesis.

Our experiments also indicate that the predictive power increases when the "useful" percentage in the dataset increases; The size of the set seems less influencing; The Amazon data set contains 53.9% useful reviews, and even though the Yelp data sets are much larger, the performance using the Amazon dataset is significantly better. Yelp datasets support this observation; The Challenge dataset has only 6.2% of "useful" reviews, while the Bay Area has 12.9%. Indeed, the performance of the classification on the Bay Area dataset is better than that for the Challenge dataset. To predict the count of useful votes per review, we use the linear regression model (LR). The evaluation metrics used are the correlation coefficient (R-squared) of the linear regression model and the Root Mean Squared Error (RMSE) as a measure of the prediction error. Table 3 details the results. The R-squared values of our models range between 0.561 and 0.871 across the Yelp and Amazon datasets; This high values of correlated R-squared suggest that there is a high correlation between the features we choose as predictors, and the count

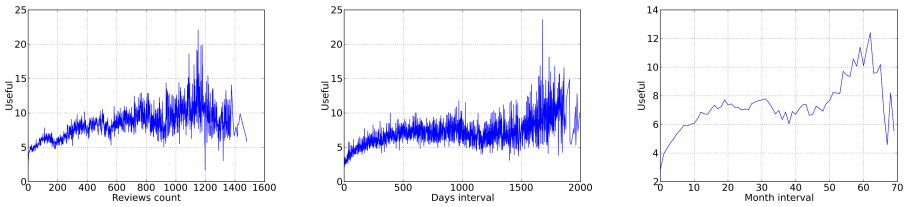
of useful votes for each review. The RMSE is supporting the accuracy of the results, the values of the RMSE are less than 1.7% of the values that the model predicts.

We proceed to check the statistical significance of each feature across all three datasets, and find contributing features. A contributing feature has a statistical significant effect on the extent by which a review is perceived useful. Specifically, we include here only features for which p$< 1\%$. For the Yelp and Yelp Challenge datasets, 20 and 18 features, respectively, are found statistically significant out of the 36 checked. For the Amazon dataset only six features are found statistically significant. Moreover, out of the four features found highly significant for all datasets, two behave differently for the Yelp datasets and for the Amazon dataset. For example, lower product rating relates to higher useful votes for the Yelp datasets, while an opposite relation exists in the Amazon dataset, i.e., higher product rating has a positive relationship with the useful vote received. Other significant results are the following: In all datasets, the higher the average useful vote a reviewer has, the higher is the number of useful votes the review receives. Similarly, the more disgust emotion is expressed, the more the review is likely to gain more useful votes over all sets. In the Yelp datasets, the number of adjectives has a negative correlation with the useful vote, while the number of punctuations and question marks has a positive relation.

5 How to Become a Local Celeb

In this Section we search for the root cause for our results. Initial explanations are obvious ones: First, it could be that experts simply shine. These reviewers manage to capture the essence of what would be important to others and explain it vividly. The second explanation is stems from an underlying mechanism in social networks - preferential attachment [12]. Accordingly, the text of reviewers who are considered influential (i.e., earn Elite badges on Yelp or high rank badges in Amazon) is up-voted because the reviewer has acquired votes previously. Still, how does one become a Yelp Elite member or an Amazon top reviewer and acquires these votes?

Up-voting a review is a measure of trust: a reader cannot up-vote a review if she does not trust the reviewer to be truthful in expressing her opinion. Voting requires the reader to decide whether she finds the review useful. A positive decision demonstrates trust, and a fundamental requirement for that trust is familiarity [8]. People are more likely to trust the familiar, and familiarity obtained through frequent exposure has the potential to engender trust [9]. A high *h-index* captures users who write a lot of useful reviews. Indeed, the data shows that reviewers with a high average are quite prolific writers. Specifically, for the *h-index* range of $\{10, 70\}$, the number of reviews written on average was between 160 and 1000. Clearly, reviewers with a high *h-index* write a lot of reviews that people tend to like. When a reviewer writes frequently she becomes familiar, and hence trusted. Do people up-vote her reviews because they are useful or because they trust her judgment? Specifically, are reviewers writing a lot of reviews more likely to receive useful votes? Fig. 3 depicts the evolution of useful votes over time for reviewers. The reviewers are grouped in bins according to their average user vote and averaged, each bin containing at least five reviewers. An interesting observation is shown in Fig. 3a. The average number of useful votes a reviewer gets grows with the number of reviews she writes. Similar results are shown in Fig. 3b and Fig. 3c: The



(a) Avg. useful per #reviews (b) Avg. useful per #days from the first review (c) Avg. useful per #months from first review

Fig. 3: Average useful votes with time/reviews# growing

average useful votes a reviewer gets grows with time passed from the first reviews she wrote. These local celebs do indeed work hard to become known, familiar and trusted online. The result of their hard work is the number of useful votes they receive.

References

1. A. Ghose and P. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 10, pp. 1498–1512, 2011.
2. S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on Empirical MNLP*, 2006, pp. 423–430.
3. Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," in *Proceedings of the 19th international conference on WWW*. ACM, 2010, pp. 691–700.
4. N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," *Electronic Commerce Research and Applications*, vol. 11, no. 3, pp. 205–217, 2012.
5. R. P. Bagozzi, M. Gopinath, and P. U. Nyer, "The role of emotions in marketing," *Journal of the Academy of Marketing Science*, vol. 27, no. 2, pp. 184–206, 1999.
6. V. Griskevicius, N. J. Goldstein, C. R. Mortensen, J. M. Sundie, R. B. Cialdini, and D. T. Kenrick, "Fear and loving in las vegas: Evolution, emotion, and persuasion," *JMR, Journal of marketing research*, vol. 46, no. 3, p. 384, 2009.
7. K. Epstude, T. Mussweiler *et al.*, "What you feel is how you compare: How comparisons influence the social induction of affect," *Emotion*, vol. 9, no. 1, p. 1, 2009.
8. D. Gefen, "E-commerce: the role of familiarity and trust," *Omega*, vol. 28, no. 6, pp. 725–737, 2000.
9. K. Siau and Z. Shen, "Building customer trust in mobile commerce," *Communications of the ACM*, vol. 46, no. 4, pp. 91–94, 2003.
10. A. Levi, O. Mokryn, C. Diot, and N. Taft, "Finding a needle in a haystack of reviews: cold start context-based hotel recommender system," in *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 2012, pp. 115–122.
11. A. Levi and O. , Mokryn, "The social aspect of voting for useful reviews," 2014. [Online]. Available: http://www.cs.mta.ac.il/~ossi/Pubs/LM2013.pdf
12. A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.