Predicting Helpful Product Reviews

Michael P. O'Mahony[†], and Pádraig Cunningham[‡] and Barry Smyth[†]

[†]CLARITY: Centre for Sensor Web Technologies,
[‡]UCD Complex and Adaptive Systems Laboratory,
School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
{michael.omahony,padraig.cunningham,barry.smyth}@ucd.ie

Abstract. Millions of users are today posting user-generated content online, expressing their opinions on all manner of goods and services, topics and social affairs. While undoubtedly useful, user-generated content presents consumers with significant challenges in terms of information overload and quality considerations. In this paper, we address these issues in the context of product reviews and present a brief survey of our work to date on predicting review *helpfulness*. In particular, the performance of a variety of different machine learning approaches is evaluated on four large-scale review datasets drawn from the TripAdvisor and Amazon domains. Our findings highlight some interesting properties of this task from a machine learning perspective and demonstrate that author reputation, the sentiment expressed in reviews and review length are among the most effective predictors of review helpfulness.

Keywords: User-generated content, product reviews, review helpfulness, classification, Amazon, TripAdvisor

1 Introduction

User-generated content comes in many forms; from the links and photographs that are shared on social networking sites such as Facebook and Myspace, to the ratings expressed for movies and books on Neflix and Amazon, to the opinions expressed on blogs and the articles contributed to Wikipedia. This content provides consumers with a wealth of information which is sourced everyday by people engaging in online activities, such as sharing data with friends, deciding which restaurant to frequent and shopping for all kinds of goods and services.

The growing volume and diversity of this content does, however, present challenges to consumers. For example, the social networking site Facebook has some 400 million active users, with each creating, on average, 70 pieces of content each month[1]. The microblogging service Twitter is also enormously popular, where approximately 55 million tweets are sent every day by almost 106 million users[2]. Thus, while the creation of user-generated content is enthusiastically

[1] http://www.facebook.com/press/info.php?statistics. Accessed 22/06/2010.
[2] http://www.businessinsider.com/twitter-stats-2010-4. Accessed:20/06/2010.

supported, the shear quantity of such content can present a significant issue for consumers when it comes to finding information that is of personal relevance. Moreover, given that virtually anybody motivated to create content is free to do so and given the pseudo-anonymous nature of the Web, there is no guarantee that posted content is reliable, balanced, insightful and free of bias. Hence, a clear need exists to assist consumers to effectively leverage this vast, but at times unreliable, information source.

In this paper, we consider these issues of quantity and quality of user-generated content in the context of product reviews. Review websites, such as Amazon, Yelp and TripAdvisor have become a key resource for shoppers, travelers and service consumers in general and assist people to make better informed purchasing decisions based on the experiences of fellow consumers. Of course, there is no guarantee that reviews are presented in a manner that is helpful to consumers or that they are independent and free of the natural bias expressed by product manufacturers or service providers. Here, we present a review of our recent work [9–12] that applies machine learning techniques to identify the most *helpful* reviews that are submitted for products. In particular, we evaluate the performance of a variety of machine learning approaches, using feature selection, on a large-scale datasets from the TripAdvisor and Amazon websites. The review features we consider are derived from the structural and readability properties of review texts, the reputation of review authors and the sentiment expressed in reviews toward the product in question[3].

Briefly, the paper is organised as follows. The feature sets used in our approach are described in Section 2. An evaluation of our approach using large collections of TripAdvisor and Amazon product reviews in presented in Section 3. We conclude by discussing the significance and applicability of our approach and presenting an outline for future work in Section 4.

2 Review Classification

Popular products often attract hundreds or even thousands of consumer reviews. Some online services, in an effort to address this form of information overload, allow users to provide feedback on the *helpfulness* of each review, and use this data to rank review lists. While this approach is welcome, many reviews — particularly the more recent ones — fail to attract any feedback. For example, in the case of the Chicago dataset (consisting of TripAdvisor hotel reviews, see Section 3.1) used in our evaluation, some 25% of reviews received no feedback at all and only 35% of reviews received feedback on 5 or more occasions. Hence the need exists for automated techniques that can predict review helpfulness in the absence of user-supplied feedback.

[3] In other work [5–8, 14, 15], review texts have also been analysed in terms of, for example, lexical (e.g. unigram and bigram distributions), syntactic (e.g. analysis of text composition in terms of the percentages of nouns, adjectives, verbs and adverbs present in the text) and semantic (e.g. positive or negative opinions expressed with respect to product features) properties. See [11] for an overview of this work.

Our objective is to distinguish between the *helpful* and *unhelpful* reviews that are submitted for products by using a supervised machine learning approach. First, we define how reviews are labeled (i.e. as either helpful or unhelpful) and then describe the set of features used to create review instances.

2.1 Establishing Ground truth

In both the TripAdvisor and Amazon domains, users can provide feedback on whether they found reviews to be helpful or not and we rely on such feedback to establish the ground truth for review helpfulness. In particular, we define the helpfulness (h_{r_i}) of a review, r_i, as $h_{r_i} = \frac{n_{i(pos)}}{n_{i(pos)}+n_{i(neg)}}$, where $n_{i(pos)}$ and $n_{i(neg)}$ are the number of positive and negative helpfulness ratings that review r_i has received, respectively. In order to distinguish the most unambiguously helpful reviews from the rest, we label a review as helpful if and only if 75% or more of the raters of that review have found it helpful, i.e. where $h_{r_i} \geq 0.75$.

2.2 Classification Features

We create review instances using features drawn from five groups and investigate how such properties influence the helpfulness of reviews as perceived by users. The features from each group are described in detail in Table 1; here we provide a brief rationale for each of the feature groups considered.

Reputation features measure the helpfulness of reviews that have previously been posted by authors, the hypothesis being that authors with an established record in terms of posting helpful reviews are likely to continue to do so in future.

Sentiment features capture the experience that the review author has in relation to the product. We note that reviews can vary in two dimensions; they can express either positive or negative sentiment in relation to the product in question and they can be perceived as being either helpful or unhelpful by other users. In our datasets (see Section 3.1), we find a strong correlation between review helpfulness and sentiment; for example, in the case of the Chicago dataset, positive sentiment (ratings of 4 or 5) was expressed in the majority (64%) of reviews, of which 59% were perceived as helpful (Figure 1). In contrast, only 33% of reviews expressing negative sentiment (ratings of ≤ 3) were perceived as helpful. Similar trends were observed for the other datasets considered.

Social features provide the degree distribution statistics in the bipartite user-product graph; for example, the number of review posted by each author and the number of reviews posted for each product. Our hypothesis is that the more prolific authors, i.e. those that have posted many reviews, tend to write more helpful reviews. We are also interested in determining whether a relationship exists between review helpfulness and the total number of reviews posted for a product.

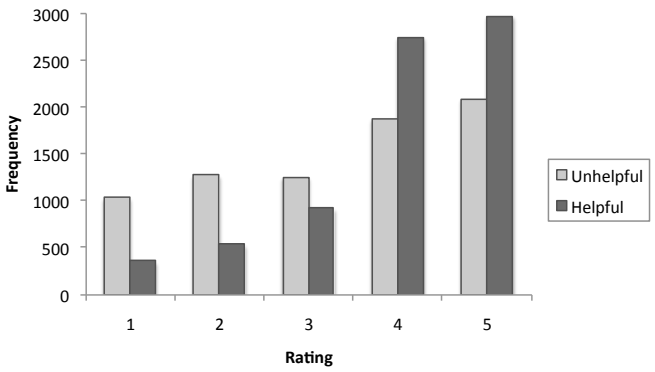

Fig. 1. Chicago dataset: relationship between sentiment (rating) and review helpfulness

Structural features provide an indication of review length, format and complexity. Review length is given by the number of words and sentences in the text and we expect longer, more comprehensive, reviews to be perceived as being more helpful by readers. Review format is captured by the number of paragraphs in the text, the percentage of alphabet and lowercase characters used. Here, we hypothesize that few paragraphs (in long reviews) or excessive use of uppercase characters (considered *shouting*) are unlikely to be appreciated by readers. The complexity of the text is indicated by the average word length (in syllables) and we hypothesize that reviews containing many long, complex words are less likely to be perceived as helpful.

Readability features measure the difficulty readers have in reading and understanding reviews [3]. These tests take into account some of the structural features described above; for example, the Gunning Fog Index is a function of the percentage of complex words in a text and the average number of words per sentence. A large number of readability tests have been proposed in the literature (see [3] for a review); in this paper, we use four widely-used readability tests (see Table 1). The rationale for considering multiple tests is that none of them correlate perfectly given the different formulations and weighting factors involved in each, and thus performance can be improved by applying a number of these tests to texts, thereby improving generalization [12]. From a review helpfulness perspective, reviews which are overly difficult (or simple) to read are unlikely to be considered helpful.

3 Evaluation

In this section, we evaluate the performance of the features described above as predictors of helpful reviews. To begin, we describe the datasets used in our evaluation and the experimental methodology employed.

Table 1. Set of features extracted from product reviews

Group	Feature	Description
Reputation	Mean author reputation	The mean helpfulness over all reviews posted by the author.
	Std. dev. author reputation	The standard deviation of helpfulness over all reviews posted by the author.
Sentiment	Author rating	The rating (on a 5-point scale) assigned by the author for the product.
	Mean author rating	The mean rating assigned by the author over all posted reviews.
	Std. dev. author rating	The standard deviation of the ratings assigned by the author over all posted reviews.
	Mean product rating	The mean rating assigned to the product over all posted reviews.
	Std. dev. product rating	The standard deviation of the ratings assigned to the product over all posted reviews.
Social	Number author reviews	The total number of reviews posted by the review author.
	Number product reviews	The total number of reviews posted for the product.
Structural	Alpha. chars. (%)	The percentage of uppercase and lowercase characters in the text.
	Uppercase chars. (%)	The percentage of uppercase characters in the text.
	'Paragraph' chars. (%)	The ratio of the number of ` ` and `<p>` tags in the text to the total number of characters in the text.
	Review length (words)	The number of words in the text.
	Number complex words	The number of *complex* words (words with 3 or more syllables) in the text.
	Review length (sentences)	The number of sentences in the text.
	Mean word length	The mean number of syllables per word.
	Mean sentence length	The mean number of words per sentence.
Readability	Flesch Reading Ease	Computed on a scale from 1 to 100, with lower scores indicating a text that is more difficult to read (e.g. a score of 30 indicates "very difficult" text and a score of 70 indicates "easy" text).
	Flesch-Kincaid Grade Level	Translates the Flesch Reading Ease score into the U.S. grade level of education required to understand a text.
	Gunning Fog Index	Estimates the number of years of formal education required for a person in order to understand a text on a first reading.
	SMOG	Acronym for Simple Measure of Gobbledygook; estimates the years of education required to completely understand a piece of text.

3.1 Datasets and Methodology

We used four large-scale review datasets for this study. We created two TripAdvisor datasets by extracting all reviews prior to April 2009 for users who had reviewed at least one hotel in either of two popular US cities, Chicago and Las Vegas. In addition, we considered two sets of Amazon product reviews [1]; these reviews relate to the *DVD* and *music* product domains.

To provide support when labeling review instances, we selected those reviews from each dataset which had received feedback (either positive of negative) on review helpfulness on a minimum of 5 occasions. In addition, we sampled from these datasets to produce balanced datasets of equal size and consisting of training data with a roughly equal representation of both helpful and unhelpful class instances. Statistics for the sampled datasets are shown in Table 2.

Table 2. Dataset statistics

Dataset	# Users	# Products	# Reviews
Chicago	6,878	6,780	15,000
Las Vegas	10,520	5,516	15,000
DVD	9,352	7,844	15,000
Music	10,427	9,496	15,000

Classification performance is evaluated using area under the ROC curve (AUC), which results in a value between 0 and 1 and is equal to the probability that a classifier will rank a randomly chosen positive instance (i.e. a helpful review) higher than a randomly chosen negative instance (i.e. an unhelpful review) [4]. However, because random guessing produces an AUC score of 0.5, no realistic classifier should have an AUC score less than 0.5. As a general rule, classifiers that achieve AUC scores of less than 0.7 are considered to provide poor performance while those that achieve AUC scores in excess of 0.9 are considered excellent [13]. Classification was performed using Weka [16] and all reported results were obtained using 10 fold cross-validation.

3.2 Classifier Performance

We compared the performance provided by four classifiers on the datasets: naïve Bayes, the J48 implementation of the C4.5 decision tree algorithm, the JRip implementation of the Ripper rule learner and a random forest learning technique. For the latter, we experimented with using 10 and 100 trees and found that both conditions provided similar results; thus we used 10 trees in our experiments.

Results are shown in Table 3 when training instances were constructed using all features. For all datasets, random forest provided the best performance, followed in turn by JRip, J48 and naïve Bayes. In terms of dataset performance, better classification was achieved for the Amazon datasets (DVD and Music) compared to the TripAdvisor datasets (Chicago and Las Vegas). Further, it can

be seen that the AUC scores achieved for both Amazon datasets were very similar over all conditions; whereas for the Amazon datasets, the scores for Las Vegas were higher than those seen for Chicago. For example, using random forrest, AUC scores of 0.799, 0.850. 0.970 and 0.969 were achieved for the Chicago, Las Vegas, DVD and Music datasets, respectively. Overall, we conclude from these findings that the feature groups described in Section 2, in conjunction with the random forrest classifier, were quite successful in predicting review helpfulness.

Table 3. Classifier comparison (AUC)

	Chicago	Las Vegas	DVD	Music
Naïve Bayes	0.709	0.739	0.868	0.871
J48	0.735	0.778	0.913	0.915
JRip	0.764	0.817	0.940	0.939
Random Forrest	0.799	0.850	0.970	0.969

We also experimented with training set size versus classification performance and the results are shown in Figure 2 using the random forrest classifier. As can be seen, AUC continues to improve for the four datasets with the addition of new data, even beyond 10,000 examples. This indicates that a complex relationship exists between review helpfulness and the predictive features and that very large training sets will be required to build good models.

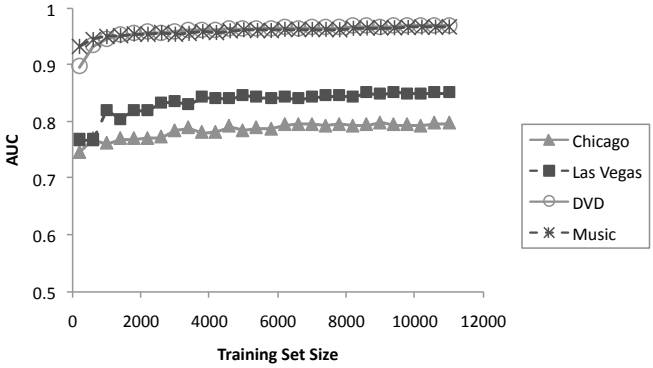

Fig. 2. AUC scores vs. training set size

3.3 Feature Groups and Feature Selection

We now consider the classification performance provided by each of the feature groups described in Section 2. Results are shown in Figure 3 using random forrest. As before, very similar results were seen for the Amazon datasets, which were significantly higher in all cases than those achieved for the TripAdvisor datasets. For all four datasets, the reputation features provided best performance (with notably very high AUC scores of 0.94 for both Amazon datasets, and between 0.7 and 0.8 for the TripAdvisor datasets), followed by sentiment features. Structural and readability features also performed quite well for the Amazon datasets, although this was not the case for the TripAdvisor datasets. For all datasets, social features performed poorly, with AUC scores of approximately 0.6 or less.

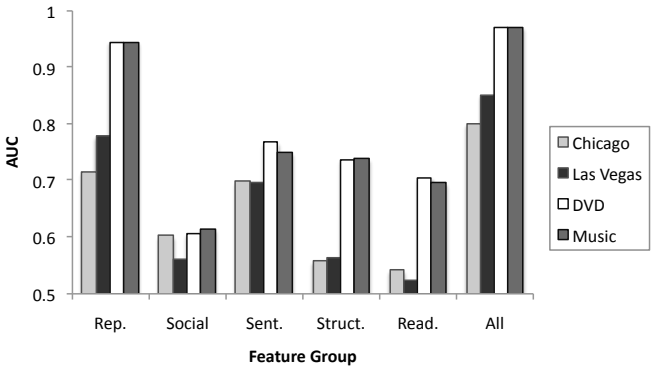

Fig. 3. AUC scores achieved by feature groups

While the above analysis examined feature groups, we now consider the performance provided by individual features using feature selection based on information gain. The subset selection strategy we use is to rank the features using information gain and then, starting with the highest scoring feature, evaluate using cross-validation the performance of a classifier built with that feature. Then add the next highest ranking feature and reevaluate; repeat until no further improvements are achieved [2]. The results of this strategy (using random forrest) are shown in Figure 4. Interestingly, random forest was able to get some useful information from all features with performance continuing to improve until all features were included. We note, however, that the inclusion of all features does result in significantly reduced performance for the J48 and (to a lesser extent) naïve Bayes classifiers. Lower-ranked features appear to be very noisy and lead to overfitting in the case of these classifiers (see [9] for details). In relation to high-ranked features, author reputation, review sentiment and review length were among the top-5 predictors of review helpfulness for all four datasets.

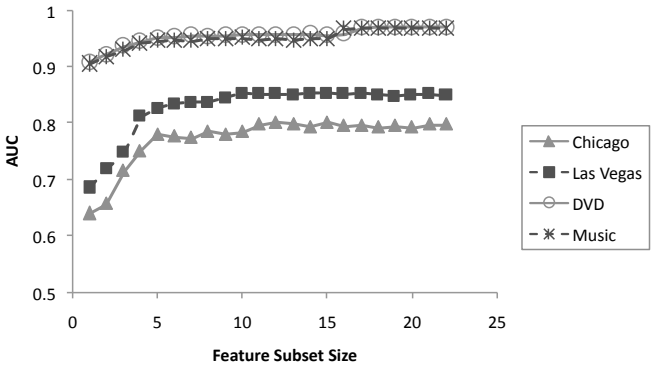

Fig. 4. AUC scores vs. feature subset size

4 Conclusions

In this paper, we have presented a brief summary of our classification-based work on predicting the helpfulness of product reviews. Using large TripAdvisior and Amazon review datasets, we have considered the performance provided by various classifiers and feature groups. The ensemble random forest classifier provided best overall performance, with reputation and sentiment features proving to be the best predictors. The best results were seen for the Amazon datasets, which were higher in all conditions evaluated than those seen for the TripAdvisor datasets.

In future work, we will analyze the differences in performance observed between the datasets drawn from the TripAdvisor and Amazon domains. In addition, we will consider addition review features (such as those proposed in related work; see footnote 3), with a view towards reducing the reliance on user reputation features, given that not all services support feedback on review helpfulness and hence such features will not always be available. In addition, we plan on applying our approach to mirco-blogging domains such as Twitter and blippr.com. From a classification perspective, reviews from these domains introduce new challenges given their short-form and noisy nature.

5 Acknowledgements

Based on work supported by Science Foundation Ireland, Grant Nos. 07/CE/I1147 and 08/SRC/I1407.

References

1. J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the*

45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), pages 440–447, Prague, Czech Republic, June 23–30 2007.
2. P. Cunningham. Dimension Reduction. In M. Cord and P. Cunningham, editors, *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, pages 91–112. Springer, 2008.
3. W. H. DuBay. The principles of readability. In *Impact Information, Costa Mesa, CA*, 2004.
4. T. Fawcett. ROC graphs: Notes and practical considerations for researchers. In *Technical Report HPL-2003-4, HP Laboratories, CA, USA*, 2004.
5. F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI 2009)*, pages 759–768, Boston, MA, USA, April 4–9 2009.
6. C.-F. Hsu, E. Khabiri, and J. Caverlee. Ranking comments on the social web. In *Proceedings of the IEEE International Conference on Computational Science and Engineering (CSE 2009)*, pages 90–97, Vancouver, Canada, 2009.
7. S.-M. Kim, P. Pantel, T. Chklovski, , and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 423–430, Sydney, Australia, July 22–23 2006.
8. Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM 2008)*, pages 443–452, Pisa, Italy, December 15–19 2008. IEEE Computer Society.
9. M. P. O'Mahony, P. Cunningham, and B. Smyth. An assessment of machine learning techniques for review recommendation. In *Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2009)*, Dublin, Ireland, August 19–21 2009.
10. M. P. O'Mahony and B. Smyth. Learning to recommend helpful hotel reviews. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys 2009)*, New York, NY, USA, October 22–25 2009.
11. M. P. O'Mahony and B. Smyth. A classification-based review recommender. *Knowledge-Based Systems*, 23:323–329, 2010.
12. M. P. O'Mahony and B. Smyth. Using readability tests to predict helpful product reviews. In *Proceedings of the 9th RIAO International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*, Paris, France, April 28–30 2010.
13. D. Streiner and J. Cairney. Whats under the ROC? An introduction to receiver operating characteristic curves. *Canadian Journal of Psychiatry*, 52(2):121–8, 2007.
14. H. Tang, S. Tan, and X. Cheng. A survey on sentiment detection of reviews. *Expert Systems With Applications*, 36(7):10760–10773, 2009.
15. W. Weerkamp and M. de Rijke. Credibility improves topical blog post retrieval. In *Proceedings of the Association for Computational Linguistics with the Human Language Technology Conference (ACL-08:HLT)*, pages 923–931, Columbus, Ohio, USA, June 16–18 2008.
16. I. H. Witten and E. Frank. *Data Mining – Practical Machine Learning Tools and Techniques, 2nd Edition*. Elsevier, 2005.