Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)

The Readability of Helpful Product Reviews

Michael P. O'Mahony and **Barry Smyth**
CLARITY: Centre for Sensor Web Technologies
UCD School of Computer Science and Informatics
University College Dublin, Ireland

Abstract

Consumers frequently rely on user-generated product reviews to guide purchasing decisions. Given the ever-increasing volume of such reviews and variations in review quality, consumers require assistance to effectively leverage this vast information source. In this paper, we examine to what extent the *readability* of reviews is a predictor of review *helpfulness*. Using a supervised classification approach, our findings indicate that readability is a useful predictor for Amazon product reviews but less so for TripAdvisor hotel reviews.

Introduction

User-generated product reviews have become a key asset to consumers, enabling assessments of product quality to be made prior to purchase. Of course, there is no guarantee that reviews are independent and free from bias or that opinions are expressed in a manner that is helpful to users. In addition, popular products often attract hundreds of consumer reviews. Thus the objective of this paper is to build on related work (Kim et al. 2006; Liu et al. 2008; O'Mahony and Smyth 2009) to develop a review classification technique that seeks to automatically identify the most *helpful* reviews from the many that are frequently submitted for products. In particular, we focus on features relating to the *readability* of review texts and examine the classification performance provided by these features.

Some online services allow users to rate 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 and hence the need for automated techniques that can reliably predict review helpfulness in the absence of such feedback.

Review Classification

In this paper, we adopt a supervised classification approach to predict review helpfulness. Using available review helpfulness feedback as the ground truth, reviews are labeled as either *helpful* or *unhelpful*. To distinguish unambiguously helpful reviews from the rest, a review is labeled helpful if and only if 75% or more of raters have found it helpful.

Prior to classication each review is translated into a feature-based instance representation. In previous work, features relating to the reputation and expertise of the review author, user sentiment toward the product, the distribution of unigrams in review texts, review length and recency were found to be useful predictors of helpful reviews (Kim et al. 2006; Liu et al. 2008; O'Mahony, Cunningham, and Smyth 2009). Here, we expand on this work to consider the performance of *readability* features on review classication.

Readability Features

Readability tests provide a means for estimating the difficulty readers have in reading and understanding text (DuBay 2004). We consider four such tests which are:

- *Flesch Reading Ease*: computes reading ease 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 US grade level of education required to understand the text.
- *Fog Index*: indicates the number of years of education required for a reader to understand the text.
- *SMOG*: indicates the years of education needed to *completely* understand a text.

Readability tests take a number of criteria into account; for example, the Fog Index is a function of the percentage of complex words (words with three or more syllables) in a text and the average number of words per sentence. See DuBay (2004) for details. From a review helpfulness perspective, we hypothesise that reviews which are too difficult to read or too simplistic are less likely to be perceived as helpful.

Evaluation

We used four large 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. We also considered two sets of Amazon reviews for *DVD* and *music* products (Blitzer, Dredze, and Pereira 2007). Similar trends were seen for the datasets drawn from each domain; thus we show results for the Chicago and DVD datasets only.

When labeling review instances, we only considered reviews which had received ≥ 5 helpfulness ratings. Further, we sampled our data to produce datasets of equal size and consisting of a roughly equal representation of helpful and unhelpful class instances. Table 1 shows dataset statistics.

Table 1: Sampled dataset statistics

Dataset	# Users	# Products	# Reviews
Chicago	6,878	6,780	15,000
DVD	9,352	7,844	15,000

Classification performance is evaluated using area under the ROC curve (AUC), which results in a value between 0 and 1 (a value of 0.5 is equivalent to random guessing). Classification was performed using a *random forest* learning technique which was found to provide good performance (O'Mahony, Cunningham, and Smyth 2009). All reported results were obtained using 10 fold cross-validation.

Classification Results

Readability features provided much better classification performance for the DVD dataset in all cases (Figure 1). The median readability values (Table 2) indicate that helpful review texts required a higher degree of reading ability on the part of the reader to understand. Wilcoxon rank sum tests indicated that all differences in medians were statistically significant at the $p < .01$ level. Greater percentages of complex words in reviews is one indicator of increased reading difficulty. The median number of complex words in helpful and unhelpful DVD reviews was 20 and 10, respectively; corresponding numbers of 20 and 14 were observed for Chicago reviews. Overall, differences between the median readability values of helpful and unhelpful reviews were greater for the DVD dataset, which correlates with the better classification performance seen for this dataset using these features.

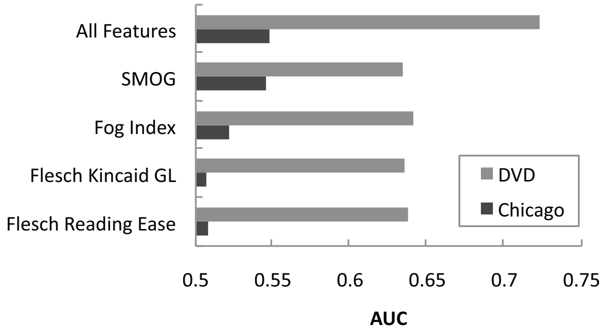

Figure 1: AUC scores achieved by readability features

Classification performance was much improved for the DVD dataset when review instances included all readability features, where an AUC score of 0.72 was achieved compared to approximately 0.64 for individual features. Given the different formulations and weighting factors involved, the readability test scores do not correlate perfectly and thus performance was improved when all features were used.

Table 2: Median readability feature values (FRE – Flesch Reading Ease, FKGL – Flesch Kincaid Grade Level)

	DVD		Chicago	
Feature	Helpful	Unhelpful	Helpful	Unhelpful
FRE	54.4	58.4	64.7	65.6
FKGL	10.8	9.8	8.3	8.1
Fog Index	13.3	12.1	10.6	10.4
SMOG	12.2	11.2	10.3	10.1

Conclusions

Although further analysis is required to understand the difference in classification performance between the Amazon and TripAdvisor datasets, in general we believe that there is merit in including readability features as part of a larger set of instance features. One advantage we envisage from using these features is the possibility of offering real-time feedback to authors when writing reviews (Bridge and Waugh 2009). For example, authors could be assisted by the system to write more helpful reviews by being warned against the use of long sentences or the excessive use of complex words. In future work, we plan on developing such a real-time feedback interface for review authors.

Acknowledgments

This work is supported by Science Foundation Ireland under Grant 07/CE/I1147.

References

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 440–447.

Bridge, D., and Waugh, A. 2009. Using experience on the read/write web: The ghostwriter system. In *Proceedings of WebCBR: The Workshop on Reasoning from Experiences on the Web (Workshop Programme of the Eighth International Conference on Case-Based Reasoning)*, 15–24.

DuBay, W. H. 2004. The principles of readability. Costa Mesa, Calif: Impact Information.

Kim, S.-M.; Pantel, P.; Chklovski, T.; ; and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 423–430.

Liu, Y.; Huang, X.; An, A.; and Yu, X. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 443–452.

O'Mahony, M. P., and Smyth, B. 2009. A classification-based review recommender. *Knowledge-Based Systems* doi:10.1016/j.knosys.2009.11.004.

O'Mahony, M. P.; Cunningham, P.; and Smyth, B. 2009. An assessment of machine learning techniques for review recommendation. In *Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science*.