<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Using readability tests to predict helpful product reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Authors(s)</strong></td>
<td>O'Mahony, Michael P.; Smyth, Barry</td>
</tr>
<tr>
<td><strong>Publication date</strong></td>
<td>2010-04-28</td>
</tr>
<tr>
<td><strong>Conference details</strong></td>
<td>Paper presented at RIAO 2010 the 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, Paris, France, April 28-30, 2010</td>
</tr>
<tr>
<td><strong>Item record/more information</strong></td>
<td><a href="http://hdl.handle.net/10197/2463">http://hdl.handle.net/10197/2463</a></td>
</tr>
</tbody>
</table>
Using Readability Tests to Predict Helpful Product Reviews

Michael P. O’Mahony
CLARITY: Centre for Sensor Web Technologies
University College Dublin, Ireland
michael.p.omahony@ucd.ie

Barry Smyth
CLARITY: Centre for Sensor Web Technologies
University College Dublin, Ireland
barry.smyth@ucd.ie

ABSTRACT
User-generated content provides online consumers with a wealth of information. Given the ever-increasing quantity of available content and the lack of quality control applied to this content, there is a clear need to enhance the user experience when it comes to effectively leveraging this vast information source. In this paper, we address these issues in the context of user-generated product reviews. We expand on recent work to consider the performance of structural and readability feature sets on the classification of helpful product reviews. Our findings, based on a large-scale evaluation of TripAdvisor and Amazon reviews, indicate that structural and readability features are useful predictors for Amazon product reviews but less so for TripAdvisor hotel reviews.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

General Terms
Algorithms, Experimentation

Keywords
User-generated product reviews, classification, helpful, TripAdvisor, Amazon

1. INTRODUCTION
The proliferation of user-generated content continues to grow in the online world. In a recent article on eMarketer.com\(^1\), it is estimated that there were 82.5 million user-generated content creators active in the US in 2008. In turn, this content was consumed by almost 116 million users. These figures clearly indicate the significance and popularity of such content and indeed the trends are set to continue, with some 115 million creators and 155 million consumers of user-generated content predicted to be active by 2013.

\(^1\)http://www.emarketer.com/Article.aspx?R=1006914

The availability of this content does, however, come at a price. Anybody motivated to create content is virtually free to do so and there is little or no quality assurance applied a priori to such content. User-generated content can therefore vary greatly in terms of quality and helpfulness; for example, it can be informative and balanced or it can be misleading and factually incorrect. In addition, the sheer quantity of available content poses significant challenges from an information overload perspective. Thus, there is a clear need to enhance the user experience by assisting users to effectively leverage this vast but at times unreliable source of content.

In this paper, we address these issues in the context of user-generated product reviews. 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. Thus the objective of this paper is to build on related work \([5, 6, 8, 9, 10]\) 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 structure and readability of review texts, and examine the classification performance provided by these features.

Briefly, the paper is organised as follows. In the next section, we describe recent work in relation to the classification of product reviews. We then describe the feature sets and classification approach that we consider in this work, which is followed by an evaluation of our approach using large collections of TripAdvisor and Amazon product reviews. We conclude by discussing the significance and applicability of our approach and an outline for future work in this area.

2. RELATED WORK
Popular products often attract hundreds of consumer reviews. Some online services 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, some 25% of hotel reviews in our Chicago dataset (see Section 4) received no feedback at all and only 35% of reviews received feedback on 5 or more occasions. Hence the need for automated techniques that can predict review helpfulness in the absence of user-supplied feedback.

In this section, we briefly describe some previous work that has been carried out in this regard.
Prior to classification, each review is translated into a feature-based instance representation. A variety of feature sets have been considered and evaluated in terms of their ability to accurately predict review helpfulness. For example, features relating to the reputation of the review author (in terms of the helpfulness of previously authored reviews) were shown in [10] to be very strong predictors of review helpfulness. In [8], reviewer expertise was shown to be a useful feature, where reviewers who were familiar with particular movie genres were more likely to produce good reviews for movies in related genres. User sentiment features, capturing how well users enjoyed their experience with a product, were also found to achieve good classification performance in [7, 10].

Further, review texts have been analysed in terms of lexical (e.g. text unigram and bigram distributions) and syntactic (e.g. analysis of text composition in terms of the percentages of nouns, adjectives, verbs and adverbs present in the text) properties. For example, Amazon product reviews were ranked according to helpfulness using SVM regression [7], with review length and unigram distribution being among the top predictive features. More recent reviews were also found to be perceived by users as being more helpful in [8].

Also of interest is related work to improve the retrieval of topical blog posts [12]. Credibility features, such as the regularity at which bloggers post, timeliness of posts, post length, spelling quality and the appropriate use of capitalisation and emoticons in the text, were considered and were found to improve retrieval performance. An approach to classify conversational and informational questions in social QA sites was proposed in [5]. Good classification performance was achieved using question category, text categorization and social network features. Further related work on review classification can be found in [6, 9, 11].

In this paper, we focus on features derived from the structural properties of review texts. Further, we analyse reviews from a readability perspective and investigate the effect of a number of readability tests on classification performance.

3. REVIEW CLASSIFICATION

Our objective is to distinguish between the helpful and unhelpful reviews that are submitted for products using a machine learning approach. For this study, we considered a large collection of reviews obtained from the TripAdvisor and Amazon domains. In both 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. To distinguish unambiguously helpful reviews from the rest, a review is labeled helpful if and only if 75% or more of review raters have found it helpful.

3.1 Structural Features

We consider a range of structural features in relation to review texts and investigate how such properties influence the helpfulness of reviews as perceived by users. In particular, the following set of features are extracted from review texts:

S1: The percentage of uppercase and lowercase characters in the text.
S2: The percentage of uppercase characters in the text.
S3: The ratio of the number of <br> and <p> tags in the text to the total number of characters in the text.
S4: The number of words in the text.
S5: The number of complex words (words with 3 or more syllables) in the text.
S6: The number of sentences in the text.
S7: The average number of syllables per word.
S8: The average number of words per sentence.

In essence, these features provide a top-level indication of review format and writing style and our hypothesis is that good format and writing style are likely to be positive indicators of helpful reviews. For example, the use of significant numbers of non-alphabet characters (e.g. emoticons) in the text may be perceived as poor writing style and thus may adversely affect review helpfulness (S1). Similarly, users are unlikely to perceive reviews authored with an excessive use of uppercase characters (or shouting) as helpful (S2).

The rationale for features S3 and S8 is that too few paragraphs (in long reviews) or over-long sentences or do not facilitate comprehension of the review text. Review length is captured by the number of words (S4) and sentences (S6) in the text, with the expectation that more comprehensive reviews are likely to be considered more helpful. An indication of the complexity of the text is given by the number of complex words used (S5) and the average number of syllables per word in the text (S7). Complex reviews, containing many words with large numbers of syllables, are less likely to be perceived as helpful (see next section).

3.2 Readability Features

As a natural complement to the above features, we also consider the use of readability tests to estimate the difficulty readers may have in reading and understanding reviews [4]. In particular, 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 into account some of the above structural parameters. For example, the Fog Index is a function of the percentage of complex words in a text and the average number of words per sentence (Eqn. 1). For further details on readability tests, refer to [4]. From a review helpfulness perspective, we hypothesise that reviews which are overly difficult to read or too simplistic are unlikely to be perceived as helpful.

\[
\text{Fog Index} = 0.4 \times \left( \frac{S4}{S6} + \frac{S5}{S4} \times 100 \right) \tag{1}
\]
4. 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.

4.1 Datasets and Methodology
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 [1]. 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 feedback on review helpfulness on ≥5 occasions. 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 the sampled dataset statistics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Users</th>
<th># Products</th>
<th># Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicago</td>
<td>6,878</td>
<td>6,780</td>
<td>15,000</td>
</tr>
<tr>
<td>DVD</td>
<td>9,352</td>
<td>7,844</td>
<td>15,000</td>
</tr>
</tbody>
</table>

Table 1: Dataset statistics

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 [9]. Reported results were obtained using 10 fold cross-validation.

4.2 Classification using Structural Features
We begin by examining the classification performance achieved by the structural features for the Chicago and DVD datasets. It is clear from the results (Figure 1) that these features provided significantly better performance for DVD reviews across all structural features.

![Figure 1: AUC scores for structural features](image)

For both datasets, the most discriminating features in terms of review helpfulness were the the number of words (S4), the number of complex words (S5) and the number of sentences (S6) in the review text. Table 2 shows the median of these feature values across helpful and unhelpful reviews. Wilcoxon rank sum tests indicated that all differences in medians were statistically significant at the p < .01 level. These findings confirm those of [7] that review length (features S4 and S6) is a useful predictor of review helpfulness, with helpful reviews being of greater median length. Examining the results for feature S4 in more detail, we can see that the ratio of the median number of words in helpful reviews to that in unhelpful reviews was 1.9 for the DVD dataset compared to 1.3 for Chicago; these differences in ratios correlate well with the classification performance provided by this feature for the DVD and Chicago datasets, where AUC scores of 0.6 and 0.57 were achieved, respectively. A similar analysis applies in respect of features S5 and S6.

### Table 2: Median structural feature values for helpful and unhelpful reviews

<table>
<thead>
<tr>
<th>Feature</th>
<th>DVD</th>
<th>Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4</td>
<td>176</td>
<td>204</td>
</tr>
<tr>
<td>S5</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>S6</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

All of the remaining structural features provided poor classification performance for Chicago dataset reviews, with AUC scores of less than 0.52 being achieved in each case (a score of 0.5 is equivalent to chance). Better performance was seen in the case of DVD reviews, however, with features such as the average number of syllables per word (S7) and the the percentage of uppercase and lowercase characters in the text (S1) achieving AUC scores of 0.63 and 0.61, respectively.

Finally we note that the best classification performance was seen for the DVD dataset when all features were used (labelled ‘All’); for example, an AUC score of 0.77 was achieved compared to 0.66 for the best performing individual feature (S5). A similar uplift in performance was not seen, however, for the Chicago dataset, where a low AUC score of 0.57 was achieved using all features.

4.3 Classification using Readability Features
As before, readability features provided much better classification performance for the DVD dataset in all cases (Figure 2). For both datasets, the median readability values (Table 3) 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. The absolute differences between the median readability values of helpful and unhelpful reviews were greater for the DVD dataset, which correlates with the better classification performance achieved for this dataset using these features.

![Figure 2: AUC scores for structural features](image)

As was observed for the structural features, classification performance was significantly improved for the DVD dataset when all readability features were used; for example, an AUC score of 0.72 was achieved using all features compared
to a score of 0.64 for the best performing individual readability test (Fog Index). This uplift in performance can be explained by the fact that the various readability test scores do not correlate perfectly given the different formulations and weighting factors involved in each. While all correlations were high (the average pairwise correlation between readability test scores was 0.9), nevertheless improved classification performance was realised by combining them. This effect (and the similar effect seen for structural features) can be understood in terms of the rationale for ensemble learning, where more than one classification model is trained on the problem and the predictions made by these models is combined in order to achieve improved generalisation [3].

5. CONCLUSIONS

Given the proliferation of user-generated content, the need for automated techniques to assist users to readily access relevant and high-quality content is clear. As one facet of this challenge, in this paper we considered the performance of structural and readability features on the classification of helpful and unhelpful product reviews. Although further analysis is required to understand the differences in classification performance achieved for Amazon and TripAdvisor reviews, in general we believe that there is merit in their inclusion as part of a larger range of classification features. One advantage these features provide is the possibility of offering real-time feedback to authors when writing reviews [2], for example, authors could be assisted to write more helpful reviews by being warned against the use of long sentences or the excessive use of complex words, and by comparing the readability of the text to that of existing helpful reviews. In future work, we plan on developing such a real-time feedback interface for review authors.

6. ACKNOWLEDGMENTS

This work is supported by Science Foundation Ireland under Grant 07/CE/11147.

7. REFERENCES