Automatically Predicting Peer-Review Helpfulness

Wenting Xiong
University of Pittsburgh
Department of Computer Science
Pittsburgh, PA, 15260
wex12@cs.pitt.edu

Diane Litman
University of Pittsburgh
Department of Computer Science &
Learning Research and Development Center
Pittsburgh, PA, 15260
litman@cs.pitt.edu

Abstract

Identifying peer-review helpfulness is an important task for improving the quality of feedback that students receive from their peers. As a first step towards enhancing existing peer-review systems with new functionality based on helpfulness detection, we examine whether standard product review analysis techniques also apply to our new context of peer reviews. In addition, we investigate the utility of incorporating additional specialized features tailored to peer review. Our preliminary results show that the structural features, review unigrams and meta-data combined are useful in modeling the helpfulness of both peer reviews and product reviews, while peer-review specific auxiliary features can further improve helpfulness prediction.

1 Introduction

Peer reviewing of student writing has been widely used in various academic fields. While existing web-based peer-review systems largely save instructors effort in setting up peer-review assignments and managing document assignment, there still remains the problem that the quality of peer reviews is often poor (Nelson and Schunn, 2009). Thus to enhance the effectiveness of existing peer-review systems, we propose to automatically predict the helpfulness of peer reviews.

In this paper, we examine prior techniques that have been used to successfully rank helpfulness for product reviews, and adapt them to the peer-review domain. In particular, we use an SVM regression algorithm to predict the helpfulness of peer reviews based on **generic** linguistic features automatically mined from peer reviews and students' papers, plus **specialized** features based on existing knowledge about peer reviews. We not only demonstrate that prior techniques from product reviews can be successfully tailored to peer reviews, but also show the importance of peer-review specific features.

2 Related Work

Prior studies of peer review in the Natural Language Processing field have not focused on helpfulness prediction, but instead have been concerned with issues such as highlighting key sentences in papers (Sandor and Vorndran, 2009), detecting important feedback features in reviews (Cho, 2008; Xiong and Litman, 2010), and adapting peer-review assignment (Garcia, 2010). However, given some similarity between peer reviews and other review types, we hypothesize that techniques used to predict review helpfulness in other domains can also be applied to peer reviews. Kim et al. (2006) used regression to predict the helpfulness ranking of product reviews based on various classes of linguistic features. Ghose and Ipeirotis (2010) further examined the socio-economic impact of product reviews using a similar approach and suggested the usefulness of subjectivity analysis. Another study (Liu et al., 2008) of movie reviews showed that helpfulness depends on reviewers' expertise, their writing style, and the timeliness of the review. Tsur and Rappoport (2009) proposed RevRank to select the most helpful book reviews in an unsupervised fashion based on review lexicons. However, studies of Amazon's product reviews also show that the per-

Class	Label	Features
Structural	STR	review length in terms of tokens, number of sentences, percentage of sentences that end with question marks, number of exclamatory sentences.
Lexical	UGR, BGR	*tf-idf* statistics of review unigrams and bigrams.
Syntactic	SYN	percentage of tokens that are nouns, verbs, verbs conjugated in the first person, adjectives / adverbs and open classes, respectively.
Semantic	TOP, posW, negW	counts of topic words, counts of positive and negative sentiment words.
Meta-data	MET	the overall ratings of papers assigned by reviewers, and the absolute difference between the rating and the average score given by all reviewers.

Table 1: Generic features motivated by related work of product reviews (Kim et al., 2006).

ceived helpfulness of a review depends not only on its review content, but also on social effects such as product qualities, and individual bias in the presence of mixed opinion distribution (Danescu-Niculescu-Mizil et al., 2009).

Nonetheless, several properties distinguish our corpus of peer reviews from other types of reviews: 1) The helpfulness of our peer reviews is directly rated using a discrete scale from one to five instead of being defined as a function of binary votes (e.g. the percentage of "helpful" votes (Kim et al., 2006)); 2) Peer reviews frequently refer to the related students' papers, thus review analysis needs to take into account paper topics; 3) Within the context of education, peer-review helpfulness often has a writing specific semantics, e.g. improving revision likelihood; 4) In general, peer-review corpora collected from classrooms are of a much smaller size compared to online product reviews. To tailor existing techniques to peer reviews, we will thus propose new specialized features to address these issues.

3 Data and Features

In this study, we use a previously annotated peer-review corpus (Nelson and Schunn, 2009; Patchan et al., 2009), collected using a freely available web-based peer-review system (Cho and Schunn, 2007) in an introductory college history class. The corpus consists of 16 papers (about six pages each) and 267 reviews (varying from twenty words to about two hundred words). Two experts (a writing instructor and a content instructor) (Patchan et al., 2009) were asked to rate the helpfulness of each peer review on a scale from one to five (Pearson correlation $r = 0.425$, $p < 0.01$). For our study, we consider the average ratings given by the two experts (which roughly follow a normal distribution) as the gold standard of review helpfulness. Two example rated peer reviews (shown verbatim) follow:

A helpful peer review of average-rating 5:

The support and explanation of the ideas could use some work. broading the explanations to include all groups could be useful. My concerns come from some of the claims that are put forth. Page 2 says that the 13th amendment ended the war. is this true? was there no more fighting or problems once this amendment was added? ...

The arguments were sorted up into paragraphs, keeping the area of interest clear, but be careful about bringing up new things at the end and then simply leaving them there without elaboration (ie black sterilization at the end of the paragraph).

An unhelpful peer review of average-rating 1:

Your paper and its main points are easy to find and to follow.

As shown in Table 1, we first mine **generic** linguistic features from reviews and papers based on the results of syntactic analysis of the texts, aiming to replicate the feature sets used by Kim et al. (2006). While structural, lexical and syntactic features are created in the same way as suggested in their paper, we adapt the semantic and meta-data features to peer reviews by converting the mentions of product properties to mentions of the history topics and by using paper ratings assigned by peers instead of product scores.[1]

[1]We used MSTParser (McDonald et al., 2005) for syntactic analysis. Topic words are automatically extracted from all stu-

In addition, the following **specialized** features are motivated by an empirical study in cognitive science (Nelson and Schunn, 2009), which suggests that students' revision likelihood is significantly correlated with certain feedback features, and by our prior work (Xiong and Litman, 2010; Xiong et al., 2010) for detecting these cognitive science constructs automatically:

Cognitive-science features (cogS): For a given review, cognitive-science constructs that are significantly correlated with review implementation likelihood are manually coded for each idea unit (Nelson and Schunn, 2009) within the review. Note, however, that peer-review helpfulness is rated for the whole review, which can include multiple idea units.[2] Therefore in our study, we calculate the distribution of *feedbackType* values (*praise*, *problem*, and *summary*) ($kappa = .92$), the percentage of problems that have *problem localization* —the presence of information indicating where the problem is localized in the related paper— ($kappa = .69$), and the percentage of problems that have a *solution* — the presence of a solution addressing the problem mentioned in the review— ($kappa = .79$) to model peer-review helpfulness. These kappa values (Nelson and Schunn, 2009) were calculated from a subset of the corpus for evaluating the reliability of human annotations[3]. Consider the example of the helpful review presented in Section 3 which was manually separated into two idea units (each presented in a separate paragraph). As both ideas are coded as *problem* with the presence of *problem localization* and *solution*, the cognitive-science features of this review are *praise%*=0, *problem%*=1, *summary%*=0, *localization%*=1, and *solution%*=1.

Lexical category features (LEX2): Ten categories of keyword lexicons developed for automatically detecting the previously manually annotated feedback types (Xiong et al., 2010). The categories are learned in a semi-supervised way based on syntactic and semantic functions, such as suggestion modal verbs (e.g. should, must, might, could, need), negations (e.g. not, don't, doesn't), positive and negative words, and so on. We first manually created a list of words that were specified as signal words for annotating *feedbackType* and *problem localization* in the coding manual; then we supplemented the list with words selected by a decision tree model learned using a Bag-of-Words representation of the peer reviews. These categories will also be helpful for reducing the feature space size as discussed below.

Localization features (LOC): Five features developed in our prior work (Xiong and Litman, 2010) for automatically identifying the manually coded *problem localization* tags, such as the percentage of problems in reviews that could be matched with a localization pattern (e.g. "on page 5", "the section about"), the percentage of sentences in which topic words exist between the subject and object, etc.

4 Experiment and Results

Following Kim et al. (2006), we train our helpfulness model using SVM regression with a radial basis function kernel provided by SVM^{light} (Joachims, 1999). We first evaluate each feature type in isolation to investigate its predictive power of peer-review helpfulness; we then examine them together in various combinations to find the most useful feature set for modeling peer-review helpfulness. Performance is evaluated in 10-fold cross validation of our 267 peer reviews by predicting the absolute helpfulness scores (with Pearson correlation coefficient r) as well as by predicting helpfulness ranking (with Spearman rank correlation coefficient r_s). Although predicted helpfulness ranking could be directly used to compare the helpfulness of a given set of reviews, predicting helpfulness rating is desirable in practice to compare helpfulness between existing reviews and new written ones without reranking all previously ranked reviews. Results are presented regarding the generic features and the specialized features respectively, with 95% confidence bounds.

4.1 Performance of Generic Features

Evaluation of the generic features is presented in Table 2, showing that all classes except syntactic (SYN) and meta-data (MET) features are sig-

dents' papers using topic signature (Lin and Hovy, 2000) software kindly provided by Annie Louis. Positive and negative sentiment words are extracted from the General Inquirer Dictionaries (http://www.wjh.harvard.edu/ inquirer/homecat.htm).

[2]Details of different granularity levels of annotation can be found in (Nelson and Schunn, 2009).

[3]These annotators are not the same experts who rated the peer-review helpfulness.

nificantly correlated with both helpfulness rating (r) and helpfulness ranking (r_s). Structural features (bolded) achieve the highest Pearson (0.60) and Spearman correlation coefficients (0.59) (although within the significant correlations, the difference among coefficients are insignificant). Note that in isolation, MET (paper ratings) are not significantly correlated with peer-review helpfulness, which is different from prior findings of product reviews (Kim et al., 2006) where product scores are significantly correlated with product-review helpfulness. However, when combined with other features, MET does appear to add value (last row). When comparing the performance between predicting helpfulness ratings versus ranking, we observe $r \approx r_s$ consistently for our peer reviews, while Kim et al. (2006) reported $r < r_s$ for product reviews.[4] Finally, we observed a similar feature redundancy effect as Kim et al. (2006) did, in that simply combining all features does not improve the model's performance. Interestingly, our best feature combination (last row) is the same as theirs. In sum our results verify our hypothesis that the effectiveness of generic features can be transferred to our peer-review domain for predicting review helpfulness.

Features	Pearson r	Spearman r_s
STR	$\mathbf{0.60 \pm 0.10}$*	$\mathbf{0.59 \pm 0.10}$*
UGR	0.53 ± 0.09*	0.54 ± 0.09*
BGR	0.58 ± 0.07*	0.57 ± 0.10*
SYN	0.36 ± 0.12	0.35 ± 0.11
TOP	0.55 ± 0.10*	0.54 ± 0.10*
posW	0.57 ± 0.13*	0.53 ± 0.12*
negW	0.49 ± 0.11*	0.46 ± 0.10*
MET	0.22 ± 0.15	0.23 ± 0.12
All-combined	0.56 ± 0.07*	0.58 ± 0.09*
STR+UGR+MET +TOP	0.61 ± 0.10*	0.61 ± 0.10*
STR+UGR+MET	$\mathbf{0.62 \pm 0.10}$*	$\mathbf{0.61 \pm 0.10}$*

Table 2: Performance evaluation of the generic features for predicting peer-review helpfulness. Significant results are marked by * ($p \leq 0.05$).

4.2 Analysis of the Specialized Features

Evaluation of the specialized features is shown in Table 3, where all features examined are significantly correlated with both helpfulness rating and ranking. When evaluated in isolation, although specialized features have weaker correlation coefficients ([0.43, 0.51]) than the best generic features, these differences are not significant, and the specialized features have the potential advantage of being theory-based. The use of features related to meaningful dimensions of writing has contributed to validity and greater acceptability in the related area of automated essay scoring (Attali and Burstein, 2006).

When combined with some generic features, the specialized features improve the model's performance in terms of both r and r_s compared to the best performance in Section 4.1 (the baseline). Though the improvement is not significant yet, we think it still interesting to investigate the potential trend to understand how specialized features capture additional information of peer-review helpfulness. Therefore, the following analysis is also presented (based on the absolute mean values), where we start from the baseline feature set, and gradually expand it by adding our new specialized features: 1) We first replace the raw lexical unigram features (UGR) with lexical category features (LEX2), which slightly improves the performance before rounding to the significant digits shown in row 5. Note that the categories not only substantially abstract lexical information from the reviews, but also carry simple syntactic and semantic information. 2) We then add one semantic class – topic words (row 6), which enhances the performance further. Semantic features did not help when working with generic lexical features in Section 4.1 (second to last row in Table 2), but they can be successfully combined with the lexical **category** features and further improve the performance as indicated here. 3) When cognitive-science and localization features are introduced, the prediction becomes even more accurate, which reaches a Pearson correlation of 0.67 and a Spearman correlation of 0.67 (Table 3, last row).

5 Discussion

Despite the difference between peer reviews and other types of reviews as discussed in Section 2, our work demonstrates that many generic linguistic features are also effective in predicting peer-review helpfulness. The model's performance can be alter-

[4]The best performing single feature type reported (Kim et al., 2006) was review unigrams: $r = 0.398$ and $r_s = 0.593$.

Features	Pearson r	Spearman r_s
cogS	0.43 ± 0.09	0.46 ± 0.07
LEX2	0.51 ± 0.11	0.50 ± 0.10
LOC	0.45 ± 0.13	0.47 ± 0.11
STR+MET+UGR (Baseline)	0.62 ± 0.10	0.61 ± 0.10
STR+MET+LEX2	0.62 ± 0.10	0.61 ± 0.09
STR+MET+LEX2+ TOP	0.65 ± 0.10	0.66 ± 0.08
STR+MET+LEX2+ TOP+cogS	0.66 ± 0.09	0.66 ± 0.08
STR+MET+LEX2+ TOP+cogS+LOC	$\mathbf{0.67 \pm 0.09}$	$\mathbf{0.67 \pm 0.08}$

Table 3: Evaluation of the model's performance (all significant) after introducing the specialized features.

natively achieved and further improved by adding auxiliary features tailored to peer reviews. These specialized features not only introduce domain expertise, but also capture linguistic information at an abstracted level, which can help avoid the risk of over-fitting. Given only 267 peer reviews in our case compared to more than ten thousand product reviews (Kim et al., 2006), this is an important consideration.

Though our absolute quantitative results are not directly comparable to the results of Kim et al. (2006), we indirectly compared them by analyzing the utility of features in isolation and combined. While STR+UGR+MET is found as the best combination of generic features for both types of reviews, the best individual feature type is different (review unigrams work best for product reviews; structural features work best for peer reviews). More importantly, meta-data, which are found to significantly affect the perceived helpfulness of product reviews (Kim et al., 2006; Danescu-Niculescu-Mizil et al., 2009), have no predictive power for peer reviews. Perhaps because the paper grades and other helpfulness ratings are not visible to the reviewers, we have less of a social dimension for predicting the helpfulness of peer reviews. We also found that SVM regression does not favor ranking over predicting helpfulness as in (Kim et al., 2006).

6 Conclusions and Future Work

The contribution of our work is three-fold: 1) Our work successfully demonstrates that techniques used in predicting product review helpfulness ranking can be effectively adapted to the domain of peer reviews, with minor modifications to the semantic and meta-data features. 2) Our qualitative comparison shows that the utility of generic features (e.g. meta-data features) in predicting review helpfulness varies between different review types. 3) We further show that prediction performance could be improved by incorporating specialized features that capture helpfulness information specific to peer reviews.

In the future, we would like to replace the manually coded peer-review specialized features (cogS) with their automatic predictions, since we have already shown in our prior work that some important cognitive-science constructs can be successfully identified automatically.[5] Also, it is interesting to observe that the average helpfulness ratings assigned by experts (used as the gold standard in this study) differ from those given by students. Prior work on this corpus has already shown that feedback features of review comments differ not only between students and experts, but also between the writing and the content experts (Patchan et al., 2009). While Patchan et al. (2009) focused on the review comments, we hypothesize that there is also a difference in perceived peer-review helpfulness. Therefore, we are planning to investigate the impact of these different helpfulness ratings on the utilities of features used in modeling peer-review helpfulness. Finally, we would like to integrate our helpfulness model into a web-based peer-review system to improve the quality of both peer reviews and paper revisions.

Acknowledgements

This work was supported by the Learning Research and Development Center at the University of Pittsburgh. We thank Melissa Patchan and Christian D. Schunn for generously providing the manually annotated peer-review corpus. We are also grateful to Christian D. Schunn, Janyce Wiebe, Joanna Drummond, and Michael Lipschultz who kindly gave us valuable feedback while writing this paper.

[5] The accuracy rate is 0.79 for predicting *feedbackType*, 0.78 for *problem localization*, and 0.81 for *solution* on the same history data set.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. In Michael Russell, editor, *The Journal of Technology, Learning and Assessment (JTLA)*, volume 4, February.

Kwangsu Cho and Christian D. Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. In *Computers and Education*, volume 48, pages 409–426.

Kwangsu Cho. 2008. Machine classification of peer comments in physics. In *Proceedings of the First International Conference on Educational Data Mining (EDM2008)*, pages 192–196.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinber g, and Lillian Lee. 2009. How opinions are received by online communities: A case study on Amazon .com helpfulness votes. In *Proceedings of WWW*, pages 141–150.

Raquel M. Crespo Garcia. 2010. Exploring document clustering techniques for personalized peer assessment in exploratory courses. In *Proceedings of Computer-Supported Peer Review in Education (CSPRED) Workshop in the Tenth International Conference on Intelligent Tutoring Systems (ITS 2010)*.

Anindya Ghose and Panagiotis G. Ipeirotis. 2010. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. In *IEEE Transactions on Knowledge and Data Engineering*, volume 99, Los Alamitos, CA, USA. IEEE Computer Society.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP2006)*, pages 423–430, Sydney, Australia, July.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, volume 1 of *COLING '00*, pages 495–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang Liu, Xiangji Guang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pages 443–452, Los Alamitos, CA, USA.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 91–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

Melissa M. Nelson and Christian D. Schunn. 2009. The nature of feedback: how different types of peer feedback affect writing performance. In *Instructional Science*, volume 37, pages 375–401.

Melissa M. Patchan, Davida Charney, and Christian D. Schunn. 2009. A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor. In *Journal of Writing Research*, volume 1, pages 124–152. University of Antwerp.

Agnes Sandor and Angela Vorndran. 2009. Detecting key sentences for automatic assistance in peer-reviewing research articles in educational sciences. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 36–44.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM2009)*, pages 36–44.

Wenting Xiong and Diane J. Litman. 2010. Identifying problem localization in peer-review feedback. In *Proceedings of Tenth International Conference on Intelligent Tutoring Systems (ITS2010)*, volume 6095, pages 429–431.

Wenting Xiong, Diane J. Litman, and Christian D. Schunn. 2010. Assessing reviewers performance based on mining problem localization in peer-review data. In *Proceedings of the Third International Conference on Educational Data Mining (EDM2010)*, pages 211–220.