Chapter 41
Metrics for Automated Review Classification: What Review Data Show

Ravi K. Yadav and Edward F. Gehringer

Abstract Peer review is only effective if reviews are of high quality. In a large class, it is unrealistic for the course staff to evaluate all reviews, so a scalable assessment mechanism is needed. In an automated system, several metrics can be calculated for each review. One of these metrics is volume, which is simply the number of distinct words used in the review. Another is tone, which can be positive (e.g., praise), negative (e.g., disapproval), or neutral. A third is content, which we divide into three subtypes: summative, advisory, and problem detection. These metrics can be used to rate reviews, either singly or in combination. This paper compares the automated metrics for hundreds of reviews from the Expertiza system with scores manually assigned by the course staff. Almost all of the automatic metrics are positively correlated with manually assigned scores, but many of the correlations are weak. Another issue is how the review rubric influences review content. A more detailed rubric draws the reviewer's attention to more characteristics of an author's work. But ultimately, the author will benefit most from advisory or problem detection review text. And filling out a long rubric may distract the reviewer from providing textual feedback to the author. The data fail to show clear evidence that this effect occurs.

Keywords Peer review systems · Rubrics · Automated metareviewing

41.1 Introduction

Users of peer review systems have a vested interest in high-quality reviewing. Students learn from both writing and receiving reviews. As a reviewer, a student learns more by delving deeply into the author's work and composing constructive

R.K. Yadav (✉) · E.F. Gehringer
Department of Computer Science, North Carolina State University, Raleigh, NC, USA
e-mail: rkyadav@ncsu.edu

E.F. Gehringer
e-mail: efg@ncsu.edu

Y. Li et al. (eds.), *State-of-the-Art and Future Directions of Smart Learning*,
Lecture Notes in Educational Technology, DOI 10.1007/978-981-287-868-7_41

criticism. As a review, an author wants to see feedback that is thoughtful and offers helpful advice for improving the work. But good reviewing is an art, and our experience demonstrates that students need to be taught to write careful reviews. Ideally, students would receive feedback on their reviewing just like peer review gives them feedback on their authoring. But, as there are many more reviews than there are works to review, no course staff has the time to give feedback on all reviews.

As part of our Expertiza project [1], we have designed an automated metareview system [2, 3] that can give a reviewer feedback on a review that (s)he is about to submit. Previous work [3] relates the outcomes of user studies on a small number of reviews from this software. This paper does not consider how students interact with the review system, but instead applies the metrics from the automated metareview system to hundreds or thousands of reviews from past classes. We report on three metrics for reviews:

- Volume, which is the number of distinct words used in the review
- Tone, which measures whether a review is more positive or negative about the work being reviewed
- Content, which classifies textual feedback within a review into one of the three categories
  - Summative, which says that some aspect of the work is good or bad
  - Problem detection, which identifies something that is wrong with the work
  - Advisory, which gives advice on improving the work

The other three metrics of our automated metareview system—coverage, relevance, and plagiarism—relate both to the review and to the work being reviewed and will not be discussed in this paper.

The results in Sect. 41.2 have to do with comparing automated versus manual metareviewing. For the last several years in the second author's courses (on computer architecture, object-oriented design, and ethics in computing), students have received grades for their reviewing. These grades have been, depending on the size of the class, determined by the instructor, or by the instructor and teaching assistants. For the sake of time, students have not been given grades on individual reviews; rather, either the instructor or a TA has read all of the reviews they wrote on other students' work on a particular assignment and assigned an "average" score for those reviews. The average score has then been multiplied by the number of reviews done to assign a grade for reviewing. Some students have done a large number of reviews, in courses where extra credit has been given for extra reviews [4]. In the vast majority of cases, though, students have been consistent in their reviewing—either consistently careful or consistently careless—so per-student grades closely approximate per-review grades.

Section 41.3 covers a much larger set of reviews from a wide variety of disciplines. It tests the hypothesis that a long rubric will "fatigue" the reviewers into not providing much formative feedback to go along with their ratings. Section 41.4 places our work in context with other works that have been done on automated analysis of reviews.

41.2 Automated Versus Manual Evaluation of Reviews

We would want our automated metareviews to give better scores to reviews that would also be scored better by a human rater.

Hypothesis Metareview metrics for review quality produced higher scores for artifacts that are rated more highly by human raters.

Using reviews done by 112 students in four classes, we computed the correspondence between grades awarded by the instructor for reviewing and parameters derived by the automated metareview algorithm. Table 41.1 gives the Pearson's correlation.

All of the correlations are fairly low, and thus, it is difficult to say that any of the characteristics detected by the automated algorithm strongly influence the grade. Positive tone is positively associated with grade, which indicates that reviews that make positive comments on the work receive on average better scores than those that do not. Negative tone is also positively correlated with grades, which means that from the standpoint of review score, it benefits a reviewer almost as much to say something negative as to say something positive. However, neutral tone is almost uncorrelated with grade; thus, comments that do not judge the work do not help a reviewer earn a higher grade. This makes sense, because a review that hesitates to take a position on the quality of some aspect of the work is less likely to be useful to the author.

With regard to content, the results are counterintuitive. Summative content is probably the least useful to an author; just saying that some characteristic is good or bad will not help the author improve the work. Problem detection is almost uncorrelated with the grade, while advisory content is weakly correlated with improved grades.

Volume is correlated with grade, as we would expect. Typically, a longer review is one that says more about the reviewed work and thus is more likely to be useful to the author. But none of the correlations are very high. There could be several reasons for this.

Table 41.1 Correlation between automated metareview parameters and staff-assigned grades

Automated metareview parameter	Pearson's correlation with grades
Tone—positive	0.341
Tone—neutral	−0.0846
Tone—negative	0.266
Summative content	0.312
Problem detection	0.0110
Advisory content	0.186
Volume	0.308

First, the Pearson's correlation is high if there is a linear correlation between two variables. But reviews are graded on a fixed scale of 0–100, while reviews can have arbitrary amounts of positive tone, advisory content, etc. Up to a point, increasing the amount of useful content can earn the reviewer a higher grade. But students who meet the instructor's expectations are likely to earn high grades, and grades for students who substantially exceed the instructor's expectations will not be much higher.

Second, we usually desire to award students some credit for effort, even if their review is not very helpful. The first year that the staff graded the students' reviews, we awarded 45 % for a review that simply filled out the score dropdowns on the review form, but did not make any textual comments on the work. After a semester or two, it became clear that this was too generous, so we lowered that to 30 %. And after the second year, we warned students that we would not award any credit for reviews that did not make at least two suggestions for how to improve the work. But most of the reviews in the study were graded before we adopted this policy. Thus, the grades given to reviews were much more compressed than the range of any of the artifacts that our metrics were looking for.

41.3 Influence of Rubric Length on Review Length

It almost goes without saying that good rubrics elicit good feedback. A rubric must be detailed enough to draw the reviewer's attention to many salient aspects of the author's work. This suggests that rubrics should be detailed. However, it seems that if the rubric is very long, students feel less of an obligation to provide textual feedback, perhaps due to weariness or out of a sense that they are giving the author enough feedback by just checking boxes and filling out dropdowns. This suggests the following:

Hypothesis Review length (text submitted by students) will tend to vary inversely with rubric length (total length of all the rubric criteria).

To investigate this hypothesis, we chose all rubrics that were used in more than 50 Expertiza reviews. This set contained 76 rubrics. The rubrics contained several kinds of criteria.

- Likert-scale rating items, each with a text box for additional comments
- Checkbox items, where the reviewer would either check a box or not
- Text areas, where a reviewer could type a long comment, e.g., to summarize a submission or write a prose review.

Figure 41.1 is a scatter plot of review length versus rubric length, each measured in characters. Note that the points in this plot fall into a discrete number of columns. This is because each review that uses a specific rubric will have the same rubric length, although the length of review done using this rubric may vary markedly.

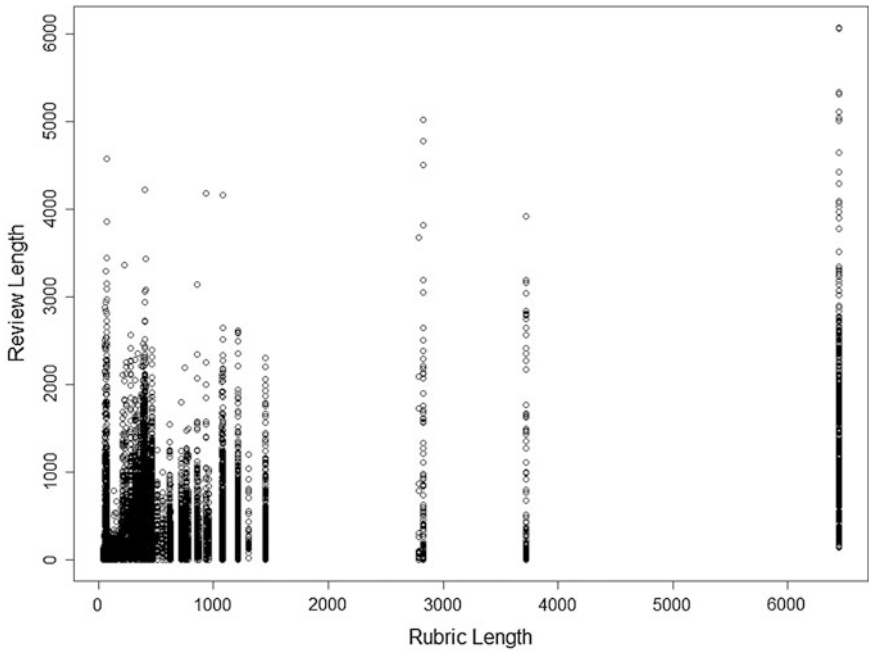

Fig. 41.1 Review length versus rubric length

From glancing at the plot, it is clear that longer rubrics are associated with longer reviews. The Pearson's correlation between review length and rubric length is 0.3225. Longer rubrics are not associated with shorter reviews. So the hypothesis is not confirmed. In fact, longer rubrics tend to induce longer responses from the students, though the effect is weak. Thus, from this sample, it seems that using a longer rubric is worthwhile.

However, if we exclude the 5 rubrics that have length >1500, a different picture emerges (Fig. 41.2). There is not an obvious correlation between review length and rubric length, and indeed, the Pearson's correlation is 0.1166. So rubric length and review length are essentially uncorrelated.

So, what's different about the longer rubrics? Well, four out of the five are used in courses in schools of education. It is not surprising that faculty involved in teaching education have a greater appreciation of the power of rubrics and that they are more motivated in encouraging their students to use them. Thus, rather than longer rubrics engendering more detailed reviews, it may just be that instructors who know how to use rubrics both write longer rubrics have their students write longer reviews.

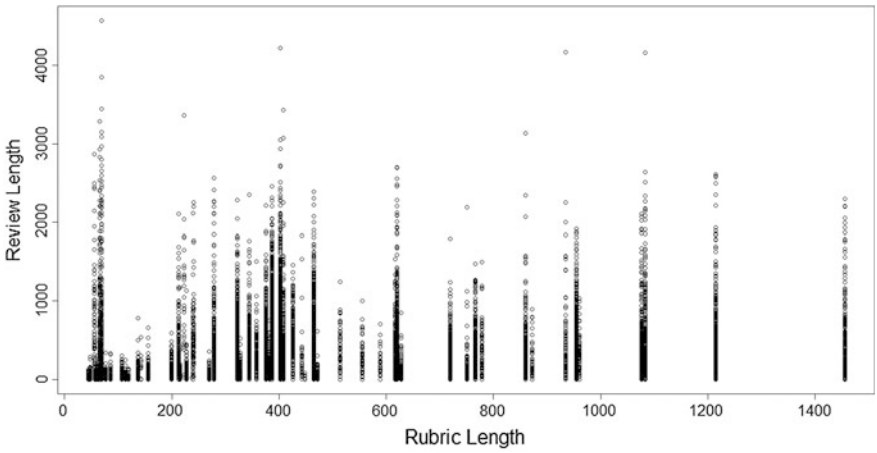

Fig. 41.2 Review length versus rubric length, for shorter rubrics

41.4 Related Work

This work, and the work cited earlier in this paper, is the first that we know of to apply multiple metrics to assessing the usefulness of peer reviews in an academic context. However, Xiong [5, 6] has used helpfulness metrics to summarize reviews in areas where the reader may be confronted with too many reviews (e.g., product reviews, hotel reviews). These build on earlier work on automatically assessing review helpfulness [7]. Similar techniques have been used for movie reviews [8]. Lu [9] attempts to rate the quality of reviewers using social network analysis.

41.5 Conclusion and Future Work

This work represents the first attempt at applying automated metareview rubrics to a large number of students' reviews. We expected that the automated metrics would give scores compatible with those assigned by human raters (in this case, the course staff). We found that the correlations between automated and manual scores were generally positive, but the degree of correlation was less than expected. It seems plausible that much of the lower-than-expected correlation comes from the fact that the metrics are linear, but we would hardly expect human-assigned grades to be uniformly distributed over the range of 0–100.

We also hypothesized that longer review rubrics might correlate with shorter reviews by students, because more of the feedback would be non-textual and because of the fatigue factor. However, we did not find such a relation. In fact, considering all review rubrics that have been used in more than 50 reviews in Expertiza, the correlation was positive, meaning that longer rubrics were associated

with longer reviews. However, on closer examination, the longest rubrics and the longest reviews come from schools of education, which are likely to place the most emphasis on use of rubrics. This evidence suggests that students may benefit from using longer rubrics, because they draw their attention to a larger number of characteristics of the work being reviewed.

This study sets the groundwork for improving automated metareviewing. It suggests the use of nonlinear metrics for comparing automated and manual reviewing. Also, metrics that found to have little correlation with review quality can be removed from automated feedback to speed it up in cases where a reviewer is waiting for automated feedback before submitting a review.

In the immediate future, we are going to work on performance tuning of our other metrics—relevance, coverage, and plagiarism—which require processing both the submission and the review, so that analysis of the submission can be performed at the time it is uploaded. We can then apply these metrics to reviews done in systems such as Mobius SLIP [10] and CrowdGrader [11], which require *authors* to rate peer reviews and attempt to devise a composite metric that comes close to predicting how highly authors will rate a review. Once such a metric is devised, it can be presented to a reviewer who is about to submit a review, along with automated advice on how to improve the review before submission. This will serve our goal of providing authors with high-quality peer reviews.

Acknowledgments This work has been supported by the U.S. National Science Foundation under grant 1432347.

References

1. Gehringer, E. F. (2009). Expertiza: Information management for collaborative learning. In: A. A. Juan Perez (Ed.), *Monitoring and assessment in online collaborative environments: Emergent computational technologies for e-learning support*. Hershey: IGI Global Press.
2. Ramachandran, L. (2013). Automated assessment of reviews. Ph.D. dissertation, North Carolina State University, May 2013.
3. Ramachandran, L., & Gehringer, E. F. Automated assessment of the quality of peer reviews using natural language processing techniques, submitted to *International Journal of Artificial Intelligence in Education.*
4. Gehringer, E. F., & Peddycord, B. W. (2013). Grading by experience points: An example from computer ethics. In: *Proceedings of Frontiers in Education 2013*, *Oklahoma City, OK, October 23–26*.
5. Xiong, W., & Litman, D. (2011). *Automatically predicting peer-review helpfulness*. Short paper presented at The 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon: Human Language Technologies (ACL-HLT).
6. Xiong, W. & Litman, D. (2014). Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In: *The 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland.*
7. Liu, Y., Huang, X., An, A., & Yu, X. (2008). Modeling and predicting the helpfulness of online reviews. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM'08*, pp. 443–452. New York: IEEE.

8. Zhuang, L., Jing, F., & Zhu, X. (2006). Movie review mining and summarization. In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp 43–50. New York City: ACM.
9. Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. (2010). Exploiting social context for review quality prediction. In: *Proceedings of the 19th International conference on World wide web* (WWW'10), pp. 691–700. New York, NY, USA: ACM. doi:10.1145/1772690.1772761
10. Palanski, M., Babik, D., & Ford, E. (2014). *Mobius SLIP: Anonymous, peer-reviewed student writing*. OBTC 2014 at Vanderbilt University.
11. de Alfaro, L., & Shavlovsky, M. (2014). CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In: *Proceedings of the 45th ACM technical symposium on Computer science education* (SIGCSE'14), pp. 415–420. New York, NY, USA: ACM. doi:10.1145/2538862.2538900, http://doi.acm.org/10.1145/2538862.2538900