

Machine Classification of Peer Comments in Physics

Kwangsue Cho
chokw@missouri.edu

School of Information Science and Learning Technologies
University of Missouri, Columbia

Abstract. As part of an ongoing project where SWORD, a Web-based reciprocal peer review system, is used to support disciplinary writing, this study reports machine learning classifications of student comments on peer writing collected in the SWORD system. The student comments on technical lab reports were first manually decomposed and coded as praise, criticism, problem detection, solution suggestion, summary, or off-task. Then TagHelper 2.0 was used to classify the codes, using three frequently used algorithms: Naïve Bayes, Support Vector Machine, and a Decision Tree. It was found that Support Vector machine performed best in terms of Cohen's Kappa.

1 Introduction

Writing is known as an important tool for learning and communication in science and engineering. However still undergraduate students have less than enough writing opportunities to learn to write. Instructors are simply overwhelmed by the workload of grading and commenting on writing assignments, and therefore tend to avoid administrating writing assignments in their courses.

As an alternative to the current expert-centric approach, peer reviewing is becoming very popular to improve this unfortunate *writing crisis* in the U.S. [1]. Peer collaboration strategy can be extremely valuable to producing more writing opportunities in the class, while it does not increase the existing workload of instructors. In addition, peer reviewing can be reliable for evaluation and effective for improving the given draft [1, 4].

Although peers are practically a good source of feedback, the helpfulness of their comments on peer writing can be very limited. One of the reasons is that undergraduate students in science and engineering are very likely to be novices. In addition, they tend to be inexperienced in writing and generating constructive comments. Therefore, it is critical for the successful use of peer reviewing that students come to generate helpful reviews for their peer writers in the absence of strong expert input. Also, reciprocal peer reviewing of writing is hard for instructor to monitor. As the class size increases, the amount of peer comments increases in an exponential way.

To address the peer review problems, in this study, I present a machine learning system that classifies student written comments as *helpful* or *not helpful*. The application takes student written comments as input and produces a classification as output. If a student's comments are classified as *not helpful*, the student reviewer is asked to revise the comments by the machine learning classifier. I report one of the three classification techniques in the system.

2 Comment Categories

There has been proliferation in feedback research. Although feedback is considered to be helpful for receivers, [6] found that 40% of the feedback intervention research they reviewed in fact revealed no positive impact or a negative impact on performance. Thus it is important to find what kinds of feedback would be helpful. Past research shows that two important features of helpful feedback are specificity and praise [2].

2.1 *Specific vs. Non-specific Comments*

One of the most important factors of helpful comments is specificity. Following Hattie and Timerley [5], helpful feedback should address three main questions: “Where am I going? (What are the goals?), How am I going (What progress is being made toward the goal?), and Where to next? (What activities need to be undertaken to make better progress?)” (p. 86). Therefore, helpful comments include what is a problem that an author should work on [problem detection] and also how the problem could be improved (solution suggestion) [2]. Straub [9] found that students prefer specific comments that provide explicit suggestions rather than vague comments and negative comments. With instructor comments, it was found that students are less likely to improve their writing if the comments are ambiguous. Consistently, Conrad and Goldstein [10] found that specific feedback from teachers tend to support student revision strategies.

2.2 *Praise vs. Criticism Comments*

Consistent with prior work showing the valued nature of praise and mitigating language in feedback, peers generally rated praise as being very helpful [2]. Some social psychologists [8] have argued that positive evaluation information is more likely to be accepted than negative information. However, affective tone is regarded as not directly improving peer writing because the tonal information does not tend to include much information about how to improve a detected problem [5]. Consistently, Kluger and DeNisi's [6] meta-analysis found praise feedback has a very small effect size. Unlike oral comments, written comments tend to make tonal information less distinct. However tonal information in written comments may still affect student writing. Gee [4] found that high school juniors who never received praise turned in shorter final drafts. Seideidman [7] found that high school students who consistently received positive feedback across eight writing assignments produced more optional rough drafts and revisions than either those who received negative comments or no comments.

3 Experiments

3.1 *Written Comment Data Collection*

Participants. This study used written comments collected from 44 undergraduate students in an intro-level Physics course who participated as a part of their course requirements. Individual students played two roles, one of writer and one of reviewer. Each student was asked to write first and revised drafts of technical research papers. In addition, each student reviewer was randomly assigned four peer papers. The reviews

were double-blinded: authors had pseudonyms and reviewers merely were identified as numbers to the authors. Their age and race information was not collected. The reciprocal peer reviewing activities were supported by the SWoRD system [1].

Writing and Reviewing Tasks. Individual students as writers wrote and revised technical lab reports about sound and human hearing. The reports consisted of title, abstract, introduction and theory, experimental setup, data analysis, conclusion, and references. To motivate students to generate a best possible quality of first drafts, they were instructed that first and revised/final drafts would be weighted equally; that is, first and final drafts would contribute equally to the final grade. Individual students as reviewers evaluated four peer drafts that were randomly selected by the SWoRD system with using a moving window algorithm. Reviewers and writers were blind to each other. Reviewers were given one week to read peer drafts and generate reviews on them. Reviewers were required to evaluate each draft both qualitatively and quantitatively along with three dimensions. For each dimension, they wrote comments and then provided a rating along a seven-point rating scale from disastrous (1) to excellent (7).

Coding Scheme of Peer Comments. All the 612 sets of written comments generated by the 44 reviewers were coded. First, each comment was decomposed into idea units which were defined as a self-contained message on a single piece of strength or weakness found in peer writing. Then, each unit was assigned to one of praise, criticism, problem detection, solution suggestion, summary, or off-task categories as discussed above. It should be noted that we rarely found criticism, summary, or off-task categories. A second coder independently evaluated 10% of randomly selected written comments. First the agreement between the two coders on segmentation was examined as the number of matched segmentations and categorizations divided by the number of total segments. For the segmentation, the coders reached 98.8 % agreement and for the categorization, they reached 98.6% agreement.

Written Comment Data Collection Procedure. All the students followed the built-in SWoRD procedures. After the instructor set the course with several parameters including the number of papers each student had to write, the number of peer reviews each paper would receive (and thus how many reviews each student had to complete), the students wrote two papers (via two drafts), each draft paper received reviews from four peer reviewers, and the students were given one week for each phase, although this research focused on the first paper. Reviewers were required to evaluate each draft, both qualitatively and quantitatively, along three dimensions, (1) Introduction, theory and experimental setup, (2) Data analysis and result, and (3) Abstract and conclusion. For each dimension, they wrote comments and then provided a rating along a seven-point rating scale from *disastrous* (1) to *excellent* (7). A rubric guided the rating task and guidelines structured the commenting-giving task.

3.2 Machine Classification Techniques of Helpful Comments

While much classification research focused on semantics, a relatively small number of studies examined non-semantic information classification in text. For the comments classification, TagHelper 2.0 [3] was used. Compared to many other text classification

tools, TagueHelper is very convenient, including a wide range of configurable text classification algorithms. To prepare the input comments, first the student comments were decomposed into idea units which were defined as a self-contained message on a single piece of strength or weakness found in peer writing. There were 2,441 idea units. Then, each unit was coded as praise, criticism, problem detection, solution suggestion, summary, or off-task comment. Second, among the total 2241 units after removing 200 units coded for summary and off-task instances were randomly split into two data sets: a training set of 1120 comment units and a test set of 1121 comment units. To develop a training model, selected were punctuation removal, unigrams, bigrams, Part-Of-Speech, line length, and rare feature removal in TagHelper. Three algorithms were used to compare its performance: Naïve Bayes, Support Vector Machines (SMO in TagHelper), and a decision tree (J48 in TagHelper). Finally, the test was used to test the performance of the models.

4 Results and Discussion

Table 1. Performance of Classification Approaches

	Training (n=1120)			Test (n=1121)		
	Naïve Bays	SVM	Decision Tree	Naïve Bays	SVM	Decision Tree
Correctly classified instances	782 (69.8%)	813 (72.6%)	729 (65.1%)	784 (66.8%)	782 (60.0%)	699 (62.4%)
Incorrectly classified instances	338 (30.2%)	307 (27.4%)	391 (34.9%)	372 (33.2%)	338 (30.0%)	421 (37.6%)
Cohen's Kappa	.52	.57	.44	.48	.54	.41

Table 1 shows the experimental results. With the training set, the highest performance measured as Cohen's Kappa was achieved with SVM. Although the performances of the models were a little decreased with the test dataset, the highest Cohen's Kappa was still found with SVM. This result is consistent with other text classification studies. To identify the source of errors or performance reduction, we analyzed the confusion matrices provided by TagHelper. All the three approaches revealed consistent problems: Praise comments were correctly categorized (80% correct vs. 20% incorrect for example in SVM) while problem detection comments tended to be confused with solution suggestion. Interestingly, problem detection was categorized as solution suggestion more than solution suggestion was categorized as problem detection.

5 Conclusion

This study has presented machine learning technologies applied for classifying peer comments in writing. As demonstrated with TagHelper, the machine learning technologies were found to be useful to categorize student comments on peer writing. Especially SVM achieved a noteworthy performance. Another important result was that the machine learning technologies, especially SVM, was good at categorizing tonal information into praise vs. non-praise. Finally, it should be noted that a hidden benefit of

the text classification technologies seems to help researchers develop coding schemes precisely.

Obviously, one of the benefits in using text classification technologies is automatically categorizing a large corpus of peer comments. This may be important in reciprocal peer reviewing of writing. Along with the class size, students tend to exchange an exponential amount of comments. Thus, automatic corpus coding technologies may be greatly helpful to help instructors to monitor reciprocal peer reviewing of writing in their classes. In addition, the current findings have been used to develop an intervention for student reviewers. Thus, based on the helpful comment model developed with TagHelper, student comments can be classified online before they are passed to their authors in order to help students generate constructive comments.

References

- [1] Cho, K., & Schunn, C. D. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, 2007, 48(3), 409-426.
- [2] Cho, K., Schunn, C., & Charney, D. Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication*, 2006, 23, 260-294.
- [3] Donmez, P., Rose, C. P., Stegmann, K., Weinberger, A., and Fischer, F. Supporting CSCL with automatic corpus analysis technology. *Proceedings of Computer Supported Collaborative Learning*, 2005.
- [4] Gee, T. Students' Responses to Teachers' Comments. *Research in the Teaching of English*, 1972, 6, 212-221.
- [5] Hattie, J., & Timperley, H. The power of feedback. *Review of Educational Research*, 2007, 77(1), 81-112.
- [6] Kluger, A. N., & DeNisi, A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 1996, 119(2), 254-284.
- [7] Seidman, E. Marking Students' Compositions: Implications of Achievement Motivation theory. *Dissertation Abstracts International*, 1968, 28, 2605-A.
- [8] Shrauger, J. S., & Schoenemann, T. J. Symbolic interactionist view of self-concept: Through the looking glass darkly. *Psychological Bulletin*, 1979, 86, 3, 549-573.
- [9] Straub, R. Students' Reactions to Teacher Comments: An exploratory study. *Research in the Teaching of English*, 1997, 31, 91-119.
- [10] Conrad, S. M., & Goldstein, L. M. ESL student revision after teacher-written comments: Text, contexts, and individuals. *Journal of Second Language Writing*, 1999, 8(2), 147-179.