Predicting Low vs. High Disparity between Peer and Expert Ratings in Peer Reviews of Physics Lab Reports

Huy V. Nguyen and Diane J. Litman

University of Pittsburgh, Pittsburgh, PA, 15260
{huynv,litman}@cs.pitt.edu

Abstract. Our interest in this work is to automatically predict whether peer ratings have high or low agreement in terms of disparity with instructor ratings, using solely features extracted from quantitative peer ratings and text-based peer comments. Experimental results suggest that our model can indeed outperform a majority baseline in predicting low versus high rating disparity. Furthermore, the reliability of both peer ratings and comments (in terms of peer disagreement) shows little correlation to disparity.

Keywords: peer review, rating disparity, peer reliability, topic models.

1 Introduction

To address instructor workload and provide students with more opportunities to develop their writing and evaluation skills, instructors are increasingly using other students in the class to review and rate student papers. Given instructor concerns about the possible low validity of peer-generated grades, research has been conducted to understand when peer grading is likely to be both reliable and valid (see [3] for a short survey). Nevertheless, from the perspective of individual students some disparity between instructor and peer grades is unavoidable. Even when there is a large positive correlation between instructors and peers (across student papers), there may still be outlier peers. Our research goal is to automatically classify peers into groups of low and high rating agreement with instructors, using only information from quantitative and qualitative peer feedback. Such a classifier could be used to better understand the validity of peer assessment, and to enhance current peer-review technology systems by flagging peer outliers whose work should be reviewed by instructors.

2 Peer Review Data

The data used in this study are peer and expert reviews of the same formal report assignment collected in Physics Lab classes at the University of Pittsburgh during 2010–2011. In each class, students were asked to describe experiments they conducted and the obtained results. For this writing task, students were required to organize their reports into sections including abstract, introduction and theory (introduction), experimental setup (experiment), data analysis and questions (analysis),

K. Yacef et al. (Eds.): AIED 2013, LNAI 7926, pp. 687–691, 2013.

Table 1. Size of datasets

Report Section	Abstract	Introduction	Experiment	Analysis	Conclusion
# Instances	362	361	362	280	362

Table 2. Means of rating disparity in the low and high groups

Mean disparity	Abstract	Introduction	Experiment	Analysis	Conclusion
Low group	0.37	0.30	0.38	0.40	0.30
High group	1.51	1.39	1.53	1.65	1.61

Left to right: reviewer, rating, comment. Nimning is an expert.
AM795712 **7** *Experiment 2's part is a little lengthy […] but everything is explained clearly. Experiment 3 and 4 were perfect.*
ATgirl **7** *Really nice job! […] I understood everything you were saying.*
dude12 **7** *A lot of equations you could probably get rid of some of the basic ones, other than that it was very good.*
sureshot58 **1** *This section was basically all equations. There was little to no theory in this section. […] Try to explain more of the symbols in each equation as many of them are unclear.*
Nimning **6** *You provide most of the critical equations which are used in this experiment. […] You are also good at balancing the equation and the description of the theory.*

Fig. 1. An example instance (for the Introduction of a student report)

and conclusion. Student reports were submitted to SWoRD [2] to be assigned randomly to peers in the class for reviewing. Student reviewers were asked to evaluate each section of the reports from their peers by providing written comments and ratings using a 7-point scale (in which 7 means excellent). The number of peers per student report varied from 1 to 7, with a mean of 2.7 [1]. In addition to peer reviewers, all classes had one or two experts review and rate each student paper; most experts were the class TAs while the others were hired graduate students. This setting makes the data ideal for our study as we can use the expert ratings as a gold standard to assess the validity of the peer ratings (in terms of agreement to expert ratings). Fig. 1 is an example of a set of ratings and comments given by 4 peers and an expert to a student report section; we use the term instance for the set of reviews for a single student report section. Because different grading rubrics were given for different report sections, we study the rating disparity between peer reviewers and experts using the 5 datasets shown in Table 1 [2], each corresponding to a report section.

3 Predicting Low vs. High Rating Disparity

Binary Classification Task. For each instance, we first compute the absolute difference between the means of the peer and expert ratings. Within each dataset, these absolute differences (**Rating Disparity**) are then used to label each instance as either **Low** (for values below the median) or **High** (for values above the median). Values equal to the

[1] Some students did not do their reviewing, while others were assigned bonus reviews.
[2] Two classes did not require an analysis section in the report, and there were some data missing, so the number of instances is not the same among sections.

median are given the label of the smaller group to make the two classes as balanced as possible. For classification, we aim to predict whether the rating disparity of an instance is Low or High. As shown in Table 2, the means of rating disparity of the high groups is higher than those of the low groups (significant in all section with $p < 0.01$). Using rating agreement between peers and experts as a proxy for peer rating validity, our model predicts low versus high validity to an extent. We however leave the measurement in [3] of validity as the target variable of prediction for future work.

Features. To develop a model for predicting binary rating disparity, we represent each instance in terms of a set of automatically computable features. First, we extract the number of peer reviewers (**#Peers**) per instance, motivated by previous findings that assigning more reviewers yields greater validity. Second, we calculate the mean (**Mn**) and standard deviation (**Std**) of the peer ratings. The mean reflects our intuition that extreme ratings are more likely to result in higher deviation from expert ratings, while the standard deviation tests whether there is a relationship between rating reliability (low standard deviation) and rating validity (low disparity). Third, as an alternative method of quantifying peer reliability, we compute topic diversity in peer comments based on topic modeling. In probabilistic topic models, documents are random mixtures over latent topics, which are represented as a probability vector whose elements are the probabilities that the document belongs to the corresponding topics. Topic diversity among documents can be measured as the distance between topic distributions using Euclidean distance (**Euc**) and Kullback–Leibler divergence[3] (**KL**). For each dataset, a standard implementation[4] of LDA [1] runs over all peer comments. Each report section forms a set of peer comments whose inter-comment topic diversity is quantified by the average distance of all comment pairs in the set.

4 Results and Discussion

Table 3 shows prediction accuracies and kappa using Weka[5] J48 decision tree algorithm to learn three models from different feature sets. Compared to the majority class baseline results, the first feature set yields significantly higher accuracies for all report sections, demonstrating that low versus high rating disparity between peers and experts is predictable using the number of peer reviews in conjunction with the rating features. Examination of the learned trees shows that the mean peer rating is the most predictive feature. The Pearson's product-moment correlation coefficients in Table 4 further show that peers and experts agree more when peers give high grades. Turning to the next feature set, the results in Table 3 show that features derived from peer comments (rather than peer ratings) also significantly outperform the majority baseline, but only for three of the five sections. However, the final columns indicate that topic features do not further improve the use of rating features alone.

[3] A non-symmetric measure of the difference between 2 probability distributions - Wikipedia

[4] GibbsLDA++: `http://gibbslda.sourceforge.net`. Number of topics is set to 50.

[5] `cs.waikato.ac.nz/ml/weka`. Experiments with logistic and SVM obtained no better results.

Table 3. Prediction performance using 10-fold cross validation

Section	Majority	#Peers + Mn + Std		#Peers + Euc + KL		All Features	
	Acc.(%)	Acc.	*K*	Acc.	*K*	Acc.	*K*
Abstract	54.98	61.66 *	0.22	56.27	0.13	61.06 *	0.21
Introduction	50.69	60.40 *	0.21	61.62 *	0.23	59.91 *	0.20
Experiment	51.10	63.15 *	0.26	58.16 *	0.15	62.82 *	0.26
Analysis	51.07	62.43 *	0.24	51.07	0.0	62.07 *	0.23
Conclusion	54.42	67.02 *	0.32	59.17 *	0.16	66.86 *	0.32

* denotes significantly better than majority baseline ($p < 0.05$). The majority baseline's kappa is 0.

Table 4. Correlation coefficients between Mn and Rating Disparity ($p < 0.01$)

Section	Abstract	Introduction	Experiment	Analysis	Conclusion
Mn	-0.21	-0.37	-0.38	-0.4	-0.35

Table 5. Correlation coefficients between topic diversity values and Std ($p < 0.01$)

Section	Abstract	Introduction	Experiment	Analysis	Conclusion
Euc	0.38	0.38	0.45	0.39	0.45
KL	0.34	0.28	0.31	0.29	0.36

Finally, to explore the relation between the reliability of peer ratings and of peer comments, the results in Table 5 show that the two topic diversity metrics both positively correlate to the standard deviation of peer ratings. However, there is no correlation between any of these features (Euc, KL, or Std) and Rating Disparity. Fig. 1 illustrates such a case: although the peer ratings have a high standard deviation of 3, the mean of 5.5 is close to the expert rating and is of low disparity.

5 Conclusion

We present preliminary results in predicting binary rating disparity between peers and experts, using only features computed from information typically available during peer review (namely, peer ratings and comments). The mean of peer ratings appears as the most predictive feature in our learned models, although topic features are also predictive in some datasets. Experimental results suggest that peer rating is likely more valid when it is high. Further, neither rating disagreement nor topic diversity (reliability) directly relates to rating disparity (validity) in our data. In the future, we hope to further improve predictive accuracy by adding features extracted from student papers themselves, and will study different rating validity measurements including the measurement used in [3], and the raw difference between peer and expert rating.

Acknowledgements. This work is supported by LRDC Internal Grants Program, University of Pittsburgh. We thank C. Schunn for providing us with the data and feedback regarding this paper. We thank the reviewers for their many helpful comments.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
2. Cho, K., Schunn, C.D.: Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. Computers and Education 48(3), 409–426 (2007)
3. Cho, K., Schunn, C.D., Wilson, R.W.: Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. Journal of Educational Psychology 98(4), 891–901 (2006)