Practical Methods for Semi-automated Peer Grading in a Classroom Setting

Zheng Yuan
zhengyuan.beihang@gmail.com
Computer Science Department, Northwestern University
Evanston, Illinois

Doug Downey
d-downey@northwestern.edu
Computer Science Department, Northwestern University
Evanston, Illinois

ABSTRACT

Peer grading, in which students grade each other's work, can provide an educational opportunity for students and reduce grading effort for instructors. A variety of methods have been proposed for synthesizing peer-assigned grades into accurate submission grades. However, when the assumptions behind these methods are not met, they may underperform a simple baseline of averaging the peer grades. We introduce SABTXT, which improves over previous work through two mechanisms. First, SABTXT uses a limited amount of historical instructor ground truth to model and correct for each peer's grading bias. Secondly, SABTXT models the thoroughness of a peer review based on its textual content, and puts more weight on the more thorough peer reviews when computing submission grades. In our experiments with over ten thousand peer reviews collected over four courses, we show that SABTXT outperforms existing approaches on our collected data, and achieves a mean squared error that is 6% lower than the strongest baseline on average.

CCS CONCEPTS

• **Applied computing → Learning management systems**.

KEYWORDS

educational technology; automated grading; peer review

ACM Reference Format:
Zheng Yuan and Doug Downey. 2020. Practical Methods for Semi-automated Peer Grading in a Classroom Setting. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340631.3394878

1 INTRODUCTION

Peer grading is a widely-used tool in classrooms, in which students are asked to review each other's submissions, and the reviews are aggregated to produce consensus assessments of each submission. Peer grading has several advantages, including reducing instructor workload [1], providing an educational opportunity for students [7, 17], and enabling more prompt feedback to students on their work [11].

A general peer review process in a classroom setting includes the following steps:

1. Submission collection: students submit individual or group submissions.
2. Review matching: a system determines which peers should review which submissions.
3. Peer feedback collection: Peers provide their numeric grades and textual comments according to certain rubrics.
4. Consensus grade estimation: a system computes submission grades based on the peer grades.
5. Instructor evaluation: To ensure the quality of peer feedback, the instructor grades a portion of the submissions, and also of the review comments. Peers are rewarded for assigning grades that are similar to the instructor's grade, and for providing helpful comments.

Since peers typically lack the subject matter mastery of the instructor, peer grades exhibit both bias and variance, which makes consensus grade estimation a challenging task. A variety of previous work [8, 14–16] has proposed peer grading methods to model peer biases and variances. However, existing methods have two limitations. First, they do not model *systematic* peer bias. That is, if most peers tend to all overestimate, or all underestimate, then the consensus grades computed by the methods will be higher or lower than the ground truth grades. Second, existing methods only consider peer grades and are not able to take advantage of textual review comments. In addition to numeric grades, peers provide textual comments that point out problems in submissions or suggestions for improvement. Comments that are thorough and high-quality could indicate that a peer review merits higher weight in the consensus grade.

In this paper, we introduce SABTXT (Semi-Automated peer Bias grading approach with TeXTual reviews), which improves peer grading accuracy by using historical instructor grades to estimate peer bias, and textual review comments to estimate review thoroughness. SABTXT models peer biases in maximum-likelihood fashion using the differences between peer grades and instructor grades in the past. SABTXT models the thoroughness of a peer review using its textual comments, and puts more weights on the more thorough peer reviews when computing submission grades. Surprisingly, we find that a simple length-based estimate of review thoroughness performs comparable to powerful supervised language models trained to match the instructor peer review quality scores (step 5, above). To the best of our knowledge, ours is the first work to explore using textual reviews to improve peer grading performance. We evaluate SABTXT on peer review data sets collected

UMAP '20, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-6861-2/20/07…$15.00
https://doi.org/10.1145/3340631.3394878

over four classes, along with three synthetic data sets. The results show that SABTXT outperforms previously proposed methods and achieves an average of 6% lower MSE (Mean Squared Error) than the strongest baseline on the classroom data.

The rest of this paper is organized as follows. In Section 2, we cover previous work in automated grading. We present SABTXT in Section 3 and evaluate it in Section 4. Section 5 concludes.

2 RELATED WORK

Peer grading has been widely used to improve learning outcome and reduce instructors's workload[2–4, 6, 12, 18]. Our work focuses on peer grading in a classroom setting. Alfaro and Shavlovsky [8] propose Vancouver algorithm, which measures each peer's grading accuracy, by comparing the grades assigned by the peer with the grades by other peers for the same submissions and gives more weight to the peer grades with higher measured accuracy. The Vancouver algorithm assumes that all peers are non-biased. However, peers are not trained graders. Thus, the unbiased peer assumption is not typically met in practice. Piech et al. [14] propose a probabilistic method to do peer grade estimation. They propose three peer grades generation models and use Gibbs sampling to do the inference. Their method estimates grader biases and reliabilities based on peer grades. Raman and Joachims [15, 16] propose methods for ordinal peer grading. They claim that besides cardinal grades [8, 14], ordinal information should be take into consideration during peer grading. They propose both ordinal and cardinal methods for peer grading. According to their results [15], their ordinal enriched method achieves better results than the probabilistic method proposed in [14].

However, all of the previous works of peer grading have two shortcomings. First, they cannot deal with systematic peer biases. For example, if most of peers overestimate submission grades, the consensus grade estimated by the peer grading algorithm will be higher than ground truth grades. Second, previous work does not take textual peer feedback into consideration. We empirically compare [8, 15] and simple average baseline with the semi-supervised method proposed in this paper. Also, we target a classroom setting, where classes have on the order of 50-100 students and the instructor feedback can cover a significant proportion of the students, rather than the large MOOC setting with much more student data (but less instructor feedback).

Finally, our use of textual peer review comments is related to work in automated grading, such as [5, 10, 13]. These work concerns grading textual content, whereas we pursue a related, but different goal of estimating the weight to assign to a peer review based on its text.

3 SABTXT APPROACH

This section first formally defines the peer grading problem and presents two mechanisms, peer bias and textual review thoroughness estimation, that are used in SABTXT.

3.1 Peer grading problem definition

We assume that, in a class, n students are given a sequence of K homeworks $HW = [hw_1, hw_2, ...hw_K]$ in total. For homework k, every student makes a submission and is given a few other students' submissions to review. We use $[s_{sid_{k1}}, s_{sid_{k2}}, ..., s_{sid_{kn}}]$ to denote the submission list. sid_{ki} is the id of the submission of student i on homework k. In peer review, students (peers) provide peer grades and textual reviews. We use the symbol pg_j^i and r_j^i to denote the peer grade and textual reviews given by peer i to submission j. For each homework, the instructor will grade a small portion of the submissions and textual reviews. Grades provided by instructors are ig_j and rg_j^i, where j is a submission and i is the peer id. The goal of peer grading is to estimate the ground truth grade, g_j, for each submission j.

3.2 Semi-supervised peer bias estimation

Since peers are not well-trained graders, peer grades may not be accurate. For example, in the data we gathered for our experiments, 61% of the peer grades are higher than the instructor grades. That is, in our data the peers are biased and tend to overestimate the grades. We also observed that if a peer overestimates submissions in the past, they are likely to overestimate in the future. Based on this observation, SABTXT models peer bias using a limited amount of historical ground truth instructor grades. This method first estimates bias for each peer by averaging the difference between the historical peer grades and the corresponding instructor grades. Then, it subtracts peer biases from peer grades and averages them as the estimated consensus grades. Formally:

$$b^i = \frac{\sum_{j' \in D} (pg_{j'}^i - ig_{j'})}{|D|} \tag{1}$$

$$\hat{g}_j = \frac{\sum_{i' \in E} (pg_j^{i'} - b^{i'})}{|E|} \tag{2}$$

Equations 1 and 2 describe how consensus grades are computed for homework k. In equation 1, D is the set of submission ids that are graded by both peer i and the instructor for homework 1 to $k-1$. b^i is the estimated bias of peer i (this is the maximum likelihood estimate assuming biases are e.g. Gaussian distributed). Equation 2 computes the estimated consensus grade of submission j from homework k. E is the peer set that reviewed submission j.

3.3 Textual review thoroughness estimation

We now present our method for using textual review comments to improve peer grading performance. Textual reviews reflect how much effort a peer spends on peer reviewing and how well the peer understands the submission. Peers who provide good textual reviews are likely to provide more accurate peer grades. Our intuition is to increase the weights of peer grades that correspond to high-quality textual reviews and down-weight the peer grades of low-quality ones. Formally, our models estimate a textual review quality r_j^i for peer i and submission j, and linearly map it to the range $[-\tau, \tau]$ as the weight w_j^i. We set $tau = 0.1$ in this work. Equation 3 computes consensus peer grades using the weights:

$$g_{n\hat{e}w} = \frac{\sum_{i \in E} (pg_j^i - b^i) * (1.0 + w_j^i)}{|E|} \tag{3}$$

$$w_j^i = \frac{len(r_j^i) - (len_{max} - len_{min})/2.0}{\frac{1.0}{\tau} * (len_{max} - len_{min})} \tag{4}$$

SABTXT uses a simple yet effective method which estimates review thoroughness using the length of textual review content. SABTXT trains a linear regression model to learn historical relation between review length and peer grading accuracy. The independent variables are $[len(r_j^i), len(r_j^i)^2, \sqrt{len(r_j^i)}]$. The dependent variable is $\frac{1.0}{(pg_j^i - ig_j)^2 + 1.0}$. To predict review thoroughness, we linearly map the range of dependent variable to $[-\tau, \tau]$. Since there is no historical data for the first homework, SABTXT linearly maps review lengths to weights as is shown in Equation 4, where len_{max} and len_{min} represent the maximum and minimum review length.

When instructors's evaluation on textual peer reviews is available, it is possible to use the evaluation to train a neural model of peer review thoroughness. We now propose two neural thoroughness estimation methods for cases when instructors evaluate reviews.

SABTXT(BERT): This approach takes a textual review as input and predicts the instructor grade on the review. After training, the neural network can predict the thoroughness of any review, including those that are not graded by the instructor. Bidirectional Encoder Representations from Transformers(BERT)[9], is a widely used method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing tasks. We fine-tune the pre-trained BERT model (Base, Uncased) using the textual reviews as the input and instructor grades as the output, in the scale of [0, 10]. We tried several loss functions, but found that a sigmoid (scaled to [0,10]) output with cross-entropy loss performed the best.

$$w_j^i = \frac{\widehat{rg}_j^i - 5.0}{1.0/\tau} \tag{5}$$

Equation 5 shows how weights w_j^i are computed given BERT-predicted scores $\widehat{rg}_j^i$.

SABTXT(BOW): BERT is a powerful model, but has over a hundred million parameters and our classroom data may be too sparse to train it effectively. Thus, we also build a smaller learned model that takes bag-of-words vectors of reviews as inputs and feeds them into a one-layer neural network (i.e. a logistic regression model) to predict instructor grades on reviews. We name this model SABTXT(BOW). Like SABTXT(BERT), SABTXT(BOW) uses Equation 5 to compute weights w_j^i.

4 EXPERIMENTS

We now present our experimental results. Our experiments aim to answer the following questions:

- How accurately can SABTXT estimate consensus grades compared to other methods?
- How much do the peer bias and textual review mechanisms of SABTXT improve peer grading performance?
- Which method for using textual reviews is most effective?
- How does SABTXT's performance vary as the amount of historical instructor grades increases?

4.1 Experimental setting

We collected four peer review data sets from an algorithm design class (EECS-336, Northwestern computer science department) from the following quarters: 2017 Spring, 2017 Fall, 2019 Spring, and 2019 Fall. The instructor assigns one to two homeworks per week in EECS-336. The students work in groups of one or two. Students submit their submissions to a course management website. About three to five peers are assigned to review each submission. Peers provide their textual reviews and grades. The instructor also grades a portion of the submissions (15%). Peers are given credit based on how close their grades are to the instructor's grades. Note that the peer assignment process makes sure that each peer has at least one assigned submission that is also graded by the instructor. The difference among EECS-336 data sets is that in the 2019 Spring and Fall, peers are asked to provide textual reviews on three aspects for each homework (including correctness of the algorithm, correctness of proofs, and clarity of writing). Then, the instructor grades a part of the textual reviews by hand. In the 2017 Fall and Spring, peers are asked to provide one overall textual review, but the instructor does not evaluate its quality (thus, for these classes we do not evaluate the neural textual review models).

To help isolate how method performance depends on specific characteristics of the peer grading distribution, we also create three synthetic data sets (syn-asymbias, syn-symbias, syn-unbias). We simulate a class with 90 students and 30 homeworks. For each homework, each student submits one submission and reviews five submissions. Ground truth grades of submissions are uniformly random sampled from $[50, 100]$. Each peer has a peer variance, which is uniformly sampled from $[2, 20]$, and a peer bias. Peer biases of syn-asymbias and syn-symbias are uniformly sampled from $[0, 20]$ and $[-10, 10]$. Peer biases of syn-unbias are zeros (unbiased). Peer grades are sampled from normal distribution: $N(peer\ bias + ground\ truth,\ peer\ variance)$. Note that the synthetic data sets do not have textual peer reviews. Table 1 brief summarises the data sets. For fair comparison, we do 2-fold cross validation on experiments that use 2019 Spring and Fall data, because SABTXT(BERT) and SABTXT(BOW) require training data. Experimental results of the rest of the data sets are evaluated using instructor ground truth.

data set	students num	homework num	review num	text reviews	text review evaluation
2019 Spring	49	14	1783	✓	✓
2019 Fall	65	14	2360	✓	✓
2017 Spring	98	17	4064	✓	×
2017 Fall	92	17	3068	✓	×
synthetic	90	30	13500	×	×

Table 1: Data sets summary

4.2 Results

Table 2 compares MSEs of SABTXT, simple average, Vancouver, and MALS (Score-Weighted Mallows) proposed in [15] on all data sets. Since synthetic data sets do not have textual peer reviews, SABTXT textual review modeling is not used for synthetic data.

The results show that SABTXT (1st column) achieves lower MSEs than the other methods for nearly all of the data sets. The one exception is that Vancouver performs the best on syn-unbias,

	SABTXT	simple ave	Vancouver	MALS
2019 Spring	**171.24**	210.04	276.93	209.83
2019 Fall	**111.15**	113.54	192.77	113.59
2017 Spring	**137.14**	141.74	192.52	139.65
2017 Fall	**227.63**	232.17	291.47	229.34
syn-asymbias	**39.83**	88.03	106.36	86.69
syn-symbias	**27.99**	38.58	46.13	37.80
syn-unbias	23.44	23.76	**22.89**	22.94

Table 2: Mean Squared Error of grade estimation for SABTXT, simple average, Vancouver and MLAS on all data sets (lower is better). SABTXT outperforms baseline methods on all data sets, except the unbiased synthetic data set syn-unbias.

	SABTXT	SABTXT	SABTXT
		(- peer bias)	(- review thoroughness)
2019 Spring	**171.24**	210.37	171.61
2019 Fall	**111.15**	111.96	112.71
2017 Spring	**137.14**	140.52	138.13
2017 Fall	**227.63**	231.55	228.44

Table 3: Ablation test: removing peer bias or textual thoroughness increases Mean Squared Error.

	SABTXT	SABTXT(BERT)	SABTXT(BOW)
2019 Spring	171.24	**166.37**	233.76
2019 Fall	**111.15**	111.88	112.71
2017 Spring	**137.14**	N/A	N/A
2017 Fall	**227.63**	N/A	N/A

Table 4: Text review thoroughness estimation methods comparison.

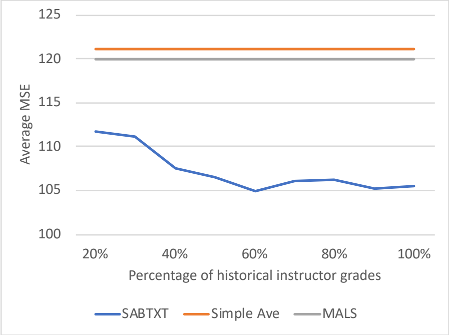

Figure 1: MSEs of SABTXT using different amount of historical instructor grades. MSE drops as historical instructor grades increases.

showing that it is most effective when data happens to be unbiased. But Vancouver's poor performance on the other data sets suggests that the unbiased assumption is too strong for real classroom data.

The MSEs of 2019 Spring and 2017 Fall are higher than the other data sets. This is due to a handful of assignments in those courses that cover newly introduced topics. In those cases, instructor grades are much lower than the peer grades, which causes higher average MSEs for all methods. This suggests we could potentially improve our methods by accounting for how bias may be higher for more challenging, unfamiliar assignments, and this is an item of future work.

To evaluate the impact of peer bias estimation and textual thoroughness estimation, we remove each from SABTXT separately and evaluate the results in Table 3. Removing either peer bias or textual thoroughness hurts performance. We found that, comparing to peer bias, the improvement of textual thoroughness estimation using review length is small.

Table 4 compares different textual review thoroughness estimation methods. Note that since the instructor did not evaluate reviews in 2017 Spring and Fall, neural model results are not available. Our results show that SABTXT(BERT) achieves the lowest MSE on 2019 Spring data comparing to SABTXT(BOW) and SABTXT. SABTXToutperforms the two neural models on 2019 Fall. We note that the length-based method is more practical when instructor evaluation of textual reviews is not available.

To test how the amount of historical instructor grades affects SABTXT, we randomly select different percentages of historical grades for peer bias estimation. We plot the average MSE over all the classroom and synthetic data in Figure 1 against the two best-performing baselines. Figure 1 shows that MSE improves given more historical instructor grades, although the improvement tapers off at about 60% of the data. Moreover, even without instructor data, SABTXT outperforms the baselines.

5 CONCLUSION

We introduce SABTXT, a semi-automated peer grading method. SABTXT improves peer grading accuracy through two mechanisms. First, by using a limited amount of historical instructor grades, SABTXT refines a model of each peer's bias throughout the course. Second, SABTXT models the thoroughness of peer reviews based on their textual content, and puts more weight on the more thorough peer reviews when estimating submission grades. The experimental results with over ten thousand peer reviews collected over four courses demonstrate that SABTXT outperforms three baseline models. Modeling peer bias is impactful, whereas review text provides a smaller improvement. We find that simple models of the text perform comparably to more powerful techniques. Our results show that our methods exhibit very different absolute performance across different classes and homeworks, which suggests that further experiments with a variety of classes beyond the four we evaluate here are necessary. In future work, we would like to continue to explore whether richer models of peers and submission content can achieve higher accuracy.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant CCF-1733860. Data used in this project is from (a) a locally modified version of Mechanical TA [18] by Madhav Suresh and Jason Hartline and (b) the PeerPal module for Canvas Learning Management System designed and built by Jacob Collins, Sadie Hood, and Jason Hartline. We also thank Eleanor O'Rourke for helpful discussion.

REFERENCES

[1] 2018. *How to Make Peer Review Successful*. https://www.thegraidenetwork.com/blog-all/2018/12/20/how-to-make-peer-reviews-successful-part-1-of-2

[2] Gabriel Badea and Elvira Popescu. 2019. Instructor Support Module in a Web-Based Peer Assessment Platform. In *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*. IEEE, 691–696.

[3] Gabriel Badea and Elvira Popescu. 2019. A Web-Based Platform for Peer Assessment in Technology Enhanced Learning: Student Module Prototype. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, Vol. 2161. IEEE, 372–374.

[4] L'hadi Bouzidi and Alain Jaillet. 2009. Can online peer assessment be trusted? *Journal of Educational Technology & Society* 12, 4 (2009), 257–268.

[5] Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 2001. Enriching Automated Essay Scoring Using Discourse Marking. (2001).

[6] Chi-Cheng Chang, Kuo-Hung Tseng, Pao-Nan Chou, and Yi-Hui Chen. 2011. Reliability and validity of Web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers & Education* 57, 1 (2011), 1306–1316.

[7] Kwangsu Cho, Christian D Schunn, and Davida Charney. 2006. Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written communication* 23, 3 (2006), 260–294.

[8] Luca de Alfaro and Michael Shavlovsky. 2013. Crowdgrader: Crowdsourcing the evaluation of homework assignments. *arXiv preprint arXiv:1308.5273* (2013).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* 1, 2 (1999), 939–944.

[11] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM conference on learning@ scale*. 75–84.

[12] Markus Mostert and Jen D Snowball. 2013. Where angels fear to tread: Online peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education* 38, 6 (2013), 674–686.

[13] Ellis Batten Page. 2003. Project Essay Grade: PEG. (2003).

[14] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579* (2013).

[15] Karthik Raman and Thorsten Joachims. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1037–1046.

[16] Karthik Raman and Thorsten Joachims. 2015. Bayesian ordinal peer grading. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. 149–156.

[17] Lloyd J Rieber. 2006. Using peer review to improve student writing in business courses. *Journal of Education for Business* 81, 6 (2006), 322–326.

[18] James R Wright, Chris Thornton, and Kevin Leyton-Brown. 2015. Mechanical TA: Partially automated high-stakes peer grading. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. 96–101.