Machine Learning for the Peer Assessment Credibility

Yingru Lin
University of Tasmania
Hobart, TAS
yingrulinn@gmail.com

Soyeon Caren Han
University of Sydney
Sydney, NSW
caren.han@sydney.edu.au

Byeong Ho Kang
University of Tasmania
Hobart, TAS
byeong.kang@utas.edu.au

ABSTRACT

The peer assessment approach is considered to be one of the best solutions for scaling both assessment and peer learning to global classrooms, such as MOOCs. However, some academic staff hesitate to use a peer assessment approach for their classes due to concerns about its credibility and reliability. The focus of our research is to detect the credibility level of each assessment performed by students during peer assessment. We found three major scopes in assessing the credibility level of evaluations, 1) Informativity, 2) Accuracy, and 3) Consistency. We collect assessments, including comments and grades provided by students during the peer assessment process and then each feedback-and-grade pair is labeled with its credibility level by Mechanical Turk evaluators. We extract relevant features from each labeled assessment and use them to build a classifier that attempts to automatically assess its level of credibility in C5.0 Decision Tree classifier. The evaluation results show that the model can be used to automatically classify peer assessments as credible or non-credible, with accuracy in the range of 88%.

CCS CONCEPTS

• **Information systems** → *Data analytics*; • **Applied computing** → *Interactive learning environments*; *E-learning*;

KEYWORDS

Peer Assessment ; Educational Data Mining ; Credibility Assessment

ACM Reference Format:
Yingru Lin, Soyeon Caren Han, and Byeong Ho Kang . 2018. Machine Learning for the Peer Assessment Credibility. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3186957

1 INTRODUCTION

MOOCs provide students video lectures with assignments, and use automated assessment, which precludes open-ended work, such as understanding and critiquing others' work. However, it is impossible to apply traditional assessment approaches to these large online classes since it requires considerable overheads in time and cost for teaching staff to assess them with detailed feedback [8]. The peer assessment approach is now considered as one of the best solutions for scaling both assessment and peer learning to the global classroom. However, some academic staff hesitate from using the peer assessment approach to their classes due to concerns about a lack of credibility and reliability [5]. One concern is that students may engage in cronyism where students come to an informal agreement to give each other the highest mark [2]. Another concern is whether students can evaluate their classmates as accurately as the instructor, who has a greater understanding of the assignment task [4]. Several credibility index and manual validation models were proposed for improving the credibility of peer assessment results but it is impossible to change the model based on different education domains [1].

Credibility in Peer Assessment The focus of our research is detecting the credibility level of each assessment performed by students during peer assessment. We found three major scopes in assessing the credibility level of evaluations, 1) Informativity, 2) Accuracy, and 3) Consistency. First, we check whether the peer assessor conforms to the communicative principles by being informative to establish coherence and continuity in the formative assessment. An assessment without informative and reasonable context would not inspire individuals nor motivate their performance, and they indicated several relevant features [6]. Secondly, the accuracy of each assessment is defined by whether the assessment feedback is aligned with the assignment marking guidelines provided by the instructor [7]. The last scope 'Consistency' can be verified with the consistency of peer assessment comment and mark. Patchan et al. [3] mentioned that students who does not take their assessments seriously provide inconsistent ratings with random judgments.

2 METHODOLOGY

We focus on assessing the credibility levels of students' assessments during the peer assessment process. The student's peer assessment data was collected from the online peer assessment system (PA system), which is implemented by the University of Tasmania (Australia). We collected the assessment data produced during semester 2, 2016 and the total number of assessment data is 13,198. The process for the online PA system can be described as follows: Teaching staff set assessment tasks and assessment criteria, allocate peer assessors for each student (assessee), and educates students in their role as an assessor. After completing the assignment task, the student submits his/her assignment to the peer assessment system. Each assessor reviews the assignment and gives an assessment score with comments supporting his/her assessment. Only the comments are available to the assessee. In return, the assessee evaluates the assessors' comments and provides a feedback score.

Next, we focus on the credibility assessment labelling. In order to train a supervised classifier over the peer assessment data, the dataset must be labeled based on the level of peer assessment credibility. We asked Mechanical Turk human evaluators to indicate credibility level for each peer assessment. Each evaluator is provided a set of a peer assessment comments provided by students

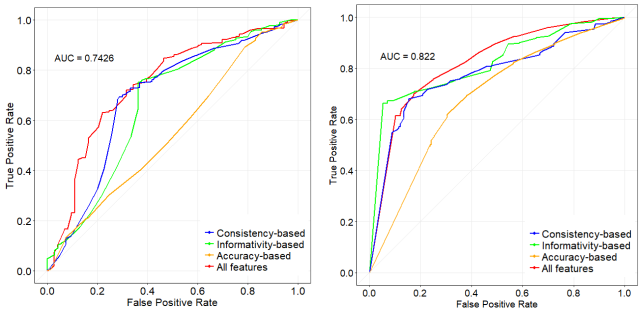
WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186957

Table 1: Results for the credibility classification

Class	TP Rate	FP Rate	Prec.	Recall	F1
Credible	0.935	0.2391	0.9066	0.9351	0.9206
Non-Credible	0.7609	0.0649	0.8253	0.7609	0.7918
W.Avg.	0.8851	0.1891	**0.8832**	**0.8851**	0.8836

and assignment marking criteria organised by teaching staff. They are asked to label the credibility level for each assessment comment based on the Likert Scale (1 to 4): 1) strongly disagree, 2) disagree , 3) agree, and 4) strongly agree. We also asked evaluators to provide a short sentence to justify their answers, and we discarded evaluations lacking that justification sentence. Each task provided one assignment marking comment, one grade and the related criteria. We randomly selected 500 cases and asked evaluators to label the level of credibility of peer assessment. After labelling, we reviewed thirteen factors that are useful to estimate the level of credibility. Thirteen factors can be classified into three major scopes in assessing the credibility level of evaluations, 1) Informativity: the amount of informative and comprehensive context delivered by students (incl. word length, character length, existence of question mark, existence of examination mark, portion of nouns, and portion of adjectives), 2) Accuracy: the degrees of the conjunction of assignment criteria and feedback comments (incl. Cosine similarity, Jaccard similarity, and Divide and Conquer distance), and 3) Consistency: the level of consistency conveyed by students (incl. positive sentiment weight, negative sentiment weight, and the percentage of grade provided by students). Finally, we trained a supervised classifier to estimate the level of credibility of students' peer assessment. To do this, we aggregate the "strongly credible" class and the "credible" class as a class "Credible", and we aggregate the "strongly non-credible" class and the "non-credible" class as a class "Non-Credible". In total, 574 cases correspond to the class 'Credible' and 526 cases correspond to the class "non-credible". We then evaluate the predictability of the credibility data. We tried a number of machine learning algorithms and the best results were achieved by a C5.0 decision tree. In the training/validation process we perform a 10-fold cross validation. As shown in Table 1. The performance for both classes is similar. The F1 in both classes is acceptable, indicating a good balance between precision and recall. The third row shows the weighted averaged performance results (88%) calculated across both classes. We also applied ROC curve for comparing performance of predicting credibility levels based on different feature groups . Figure 1 shows ROC curves comparing the performance with C5.0 Decision Tree and K-Nearest Neighbor at predicting credibility levels given three different feature groups: consistency-based feature group, informativity-based feature group, accuracy-based feature group and all three groups. we can see that AUCs in case of all features are the highest indicating that using all features have the highest accuracy rate. If all features are used, it takes multiple aspects into consideration at the same time, which gives a more uniformed result.



(a) ROC- C5.0 Decision Tree **(b) ROC- K-Nearest Neighbour**

Figure 1: ROC curve (with C5.0 Decision Tree and K-Nearest Neighbour) at predicting future class participation

3 CONCLUSIONS

The main objective of this research is to determine if we can automatically assess the level of credibility of assessment performed by students during the peer assessment process. The evaluation results show that the model can be used to automatically classify peer assessments in different domains as credible or non-credible, with accuracy in the range of 88%. Compared to other peer assessment credibility validation models, the proposed model can be used and updated in different education domains. We believe this paper is a first study of how a machine learning technique can be used for assessing the level of credibility in peer assessment. For future work, we plan to extend the experiments to larger datasets and explore more deeply the other factors that may lead students to declare an assessment as credible.

ACKNOWLEDGMENTS

This work was supported by the Industrial Core Technology Development Program (10049079, Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea). This work was also funded by the Asian Office of Aerospace Research and Development(AOARD) grant, #FA2386-16-1-404

REFERENCES

[1] Burford Furman and William Robinson. 2003. Improving engineering report writing with calibrated peer review/sup TM. In *Frontiers in Education, 2003. FIE 2003 33rd Annual*, Vol. 2. IEEE, F3E_14–F3E_16.
[2] Douglas Magin. 2001. Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education* 26, 1 (2001), 53–63.
[3] Melissa M Patchan, Christian D Schunn, and Russell J Clark. 2017. Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education* (2017), 1–16.
[4] Daniel Reinholz. 2016. The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education* 41, 2 (2016), 301–315.
[5] Sally S. Richmond, Kailasam Satyamurthy, and Joanna F. DeFranco. 2016. Exploring the Value of Peer Assessment. *American Society for Engineering Education* (2016).
[6] Yang Song, Zhewei Hu, Edward F Gehringer, Julia Morris, Jennifer Kidd, and Stacie Ringleb. 2016. Toward Better Training in Peer Assessment: Does Calibration Help?. In *EDM (Workshops)*.
[7] Hoi K Suen. 2014. Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning* 15, 3 (2014).
[8] Seng Yue Wong, Wee Jing Tee, and Wei Wei Goh. 2016. A Comparative Analysis Between Teacher Assessment and Peer Assessment in Online Assessment Environment for Foundation Students. In *Assessment for Learning Within and Beyond the Classroom*. Springer, 381–389.