Less But Enough: Evaluation of peer reviews through pseudo-labeling with less annotated data

Chengyuan Liu
North Carolina State University
cliu32@ncsu.edu

Divyang Doshi
North Carolina State University
ddoshi2@ncsu.edu

Ruixuan Shang
University of North Carolina at Chapel Hill
rshang@unc.edu

Jialin Cui
North Carolina State University
jcui9@ncsu.edu

Qinjin Jia
North Carolina State University
qjia3@ncsu.edu

Edward Gehringer
North Carolina State University
efg@ncsu.edu

A peer-assessment system provides a structured learning process for students and allows them to write textual feedback on each other's assignments and projects. This helps instructors or teaching assistants perform a more comprehensive evaluation of students' work. However, the contribution of peer assessment to students' learning relies heavily on the quality of the review. Therefore, a thorough evaluation of the quality of peer assessment is essential to assuring that the process will benefit students' learning. Previous studies have focused on applying machine learning to evaluate peer assessment by identifying characteristics of reviews (e.g., Do they mention a problem, make a suggestion, or tell the students where to make a change?). Unfortunately, collecting ground-truth labels of the characteristics is an arbitrary, subjective, and labor-intensive task. Besides in most cases, those labels are assigned by students, not all of whom are reliable as a source of labeling. In this study, we propose a semi-supervised pseudo-labeling approach to build a robust peer assessment evaluation system to utilize large unlabeled datasets along with only a small amount of labeled data. We aim to evaluate the peer assessment from two angles: Detect a problem statement (Does the reviewer mention a problem with the work?) and suggestion (Does the reviewer give a suggestion to the author?)

Keywords: peer assessment evaluation, semi-supervised learning, pseudo labeling, problem statement detection, suggestion detection.

1. Introduction

Peer assessment has long been used as a pedagogical technique in project-based courses (Li et al. 2020, Lundstrom and Baker 2009, Topping 1998, Topping 2009). Peer assessment in an online system typically allows students to provide numerical scores as well as give textual feedback on other teams' work. It has been shown remarkably effective for improving students' learning and teaming skills (Li et al., 2020). Peer assessment can also help instructors evaluate student work and assign grades more holistically. Double et al. (2020) presented a meta-analysis suggesting that peer assessment improves academic performance even more than teacher assessment. However, the reliability and validity of peer assessment rely completely upon review quality (van Zundert et al., 2010). High-quality reviews can help authors precisely identify issues in their work and make corresponding revisions. Low-quality reviews could be unhelpful or even detrimental to students' learning.

Hence, there is a growing interest among peer assessment researchers in evaluating review quality (Nelson and Schunn, 2009). However, having the instructors or teaching assistants to evaluate or grade all peer-review comments would be extremely time-consuming. Consequently, several studies have investigated machine-based automated review evaluation with the help of natural language processing techniques as well as machine learning algorithms. Nelson and Schunn (2009) carried out a pioneering study on identifying high-quality reviews by investigating the features in the textual review comments and determining what type of comments are most helpful and why. Xiao et al. proposed a machine-learning NLP-based approach for finding problem statements (Xiao et al., 2020) (e.g., Do they mention any problems in the work that required revisions) and suggestions (e.g., Do they provide any suggested solutions on how to revise the work) (Zingle et al., 2019) in the review comments.

As with most AI tasks, the biggest challenge for applying machine learning or deep-learning algorithms on peer assessment evaluation is collecting a good amount of high-quality labeled data (Xiao et al., 2020). Identifying whether review comments contain problem statements and suggestions is sometimes arbitrary and subjective; furthermore, the same review will commonly be labeled differently by student taggers. This creates a formidable obstacle in collecting precise and reliable labels for the peer-review analysis. Researchers have suggested approaches to tackling this unreliable and insufficient labeling issue. Jia et al. (2021) carried out an annotation process with two graduate students as the tagging experts and measured the inter-annotator agreement between them to gurantee the reliability of labeling. Xiao et al. (2020) proposed to apply transfer learning and active learning to tackle the insufficient-labeling problem. Transfer learning uses knowledge from a relevant task that had already been learned on a huge labeled dataset and fine-tuned on the downstream tasks. Meanwhile, active learning is also effective by helping to select the most representative samples (which are hard for the model to annotate) from the unlabeled dataset then pass to the domain experts to annotate and then progressively enlarging the training labeled dataset. However, those approaches are either based on out-domain knowledge or human intervention. None of the past works investigated utilizing only the data itself to improve the model performance, and in particular, make good use of review comments with unreliable labels, and treat them as unlabeled data. Fortunately, semi-supervised learning has been proven to be an effective way to address these issues (Goldberg, 2009). Semi-supervised learning can learn with labeled as well as unlabeled data, and reduces the need for human annotators like in active learning, but is still able to achieve good performance.

Semi-supervised learning is a learning paradigm that stands between unsupervised and supervised learning (Goldberg, 2009). The goal of semi-supervised classification is to train a classifier that uses both a small labeled dataset and a large unlabeled dataset to achieve better performance than a traditional supervised classifier trained only on the labeled data. Basic approaches to semi-supervised learning involve a well-known technique called pseudo-labeling (Lee et al., 2013), in which a classification model is first trained on the labeled dataset and then used to infer pseudo-labels on the unlabeled dataset. Then the unlabeled dataset with pseudo-labels will be combined with the labeled dataset, so that the predicted labels are used as ground truth. This allows us to progressively scale up the labeled dataset in order to train a more robust classification model.

Xie et al. (2020) conducted an extensive study on pseudo-labeling, using a self-training method with "student-teacher" models on image classification tasks. This work achieved an outstanding result. The idea is very similar to the pseudo-labeling approach. Initially a "teacher" model is trained on a small labeled dataset and then used to predict pseudo-labels on the unlabeled dataset, whereupon a "student" model will be trained on the combination of the labeled and the pseudo-labeled dataset; these steps will be run iteratively. In each iteration, once the "student" model is trained, it will be used as the new "teacher" model to generate predictions in the next iteration. This has proven to be a promising approach, progressively collecting more labeled data to address the data-insufficiency issue. Pseudo-labeling and self-training strategies have been widely applied in computer-vision tasks. (Xie et al. 2020, Nye et al. 2021) also demonstrate the power of semi-supervised learning in determining engaged and disengaged behavior in the educational context. Their approach is slightly different, utilizing K-means clustering to predict initial pseudo-labels and then using the classifier with a semi-supervised approach. This work showed that semi-supervised learning can also be applied quite effectively in the educational domain. However, very few studies have used this approach in analyzing peer-review textual comments. This paper applies natural language processing and text-classification models to apply pseudo-labeling to improve the detection of characteristics in peer review comments.

Liu et al. (2022) carried out the preliminary work of this study to apply pseudo-labeling to improve the performance of detecting problem-statement in peer-review. As an enhancement, this study conducted extensive experiment on the detection of both the suggestion and problem-statement. We aimed to provide a more comprehensive assessment of peer-review from multi-perspectives, and examined the performance of pseudo-labeling in multiple tasks. In addition, this study investigated muti-task learning techniques to target two detection tasks simultaneously during the model training phase.

The main pedagogical contribution of this study is to show how to deploy our student taggers more effectively and eventually build a robust auto-labeling system. More specifically, in our peer-assessment system, the labels can only be collected from student taggers, and if they are requested to label a large number of review comments, they may potentially become careless, resulting in poor labeling quality. We could deploy them more successfully by applying the pseudo-labeling approach to build a text classifier for evaluating peer reviews with considerably less labeled data.

2. Methodology

2.1. Automated peer-review quality evaluation

Although peer review is widely accepted in educational settings, the effectiveness of peer review in promoting students' learning can vary significantly. Most research has investigated the overall pedagogical contribution of peer review of writing. However, research on evaluating review quality is particularly lacking.

Nelson and Schunn (2009) demonstrated that high-quality reviewing has proven to be very beneficial to students' learning. That paper proposed an approach to determining what type of feedback is most helpful and why it is helpful to students' writing performance. The authors also listed the features for identifying high-quality peer reviews. This study laid an excellent foundation for later research projects on automatically detecting characteristics of peer-review comments.

The earliest study on AI-enabled peer-review quality evaluation was conducted by Cho (2008). This study proposed a machine-learning algorithm to evaluate peer reviews collected from SWoRD—a web-based reciprocal peer-review system. The review data was encoded for multiple characteristics such as problem detection, solution suggestion, etc., and then several traditional machine learning algorithms (Naïve Bayes, SVM, and Decision Tree) were applied to the text-classification task to evaluate quality.

Subsequently, automated evaluation became increasingly fashionable in peer assessment. Xiong et al. applied supervised machine learning to automatically identify problem localization (pinpoint the location of where the problem is) (Xiong and Litman 2010, Xiong et al. 2010) and helpfulness (Xiong and Litman, 2011) in peer review comments using NLP techniques. Zingle et al. (2019) describe a method for automatically detecting suggestions in review text, Xiao et al. (2020) proposed to auto-detect problem-statement in review comments.

Our study introduces an intriguing approach for automatically assessing review quality by detecting problem statements and suggestions in the comments and applying a semi-supervised learning approach to address the problem of labeled-data insufficiency. Our goal is to use less high-quality data to guide the model training and apply the pseudo-labeling approach to make label inferences and further scale up the labeled data to achieve better model performance. Pedagogically, this work can help students get instant and accurate feedback on the reviews they write and enable them to improve their reviewing as well as significantly reduce the workload of the student taggers who label those characteristics in peer-review comments.

2.2. Semi-supervised Learning & Pseudo-labeling

Deep learning has achieved great success in the area of artificial intelligence; however, most of the state-of-the-art (SotA) models were trained using supervision, which required a large labeled dataset to attain excellent performance (Goldberg, 2009). In most cases, data labeling was a labor-intensive and time-consuming task; Besides, even if we devote the time to do this, we would still be ignoring potential insights from the unlabeled dataset, which is far easier to collect in the real world. Semi-supervised learning has shown promise by using both labeled and unlabeled data. The objective of semi-supervised learning is to improve learning behavior by combining labeled and unlabeled data, or equivalently, to achieve the same model performance with a relatively small labeled dataset.

Pseudo-labeling is one of the most effective and efficient methods in semi-supervised learning (Lee et al., 2013). With pseudo-labeling, the initial model is trained on the labeled dataset:

$$D_L = \left\{(x^i, y^i)\right\}_{i=1}^{N_L} \tag{1}$$

where x^i represents each input, $y^i \subseteq \{0, 1\}$ are the corresponding labels where 0 represents "the review does not include a problem statement / suggestion" and 1 represents "the review does include a problem statement / suggestion", N_L is the size of the labeled dataset. There is also an unlabeled dataset:

$$D_U = \left\{(x^i)\right\}_{i=1}^{N_U} \tag{2}$$

where x^i represents each input without labels, and N_U is the size of the unlabeled dataset. In most cases $N_U \gg N_L$, so we believe that the unlabeled dataset may potentially contain more valuable features than the labeled set. Next, the model trained in the initial step generates predictions $\tilde{y}^i$ as pseudo-labels on the unlabeled set D_U; hence we can construct a pseudo-labeled set:

$$D_U = \left\{(x^i, \tilde{y}^i)\right\}_{i=1}^{N_U} \tag{3}$$

where $\tilde{y}^i$ will be used as the ground-truth label y^i to compute the loss in back-propagation in the next training phase, after this, another model is trained on the combination dataset:

$$D_C = D_L + D_U \tag{4}$$

and we believe that with more training data, the model will become still better.

Pseudo-labeling is often performed through an iterative process rather than one-step label generation. Self-training (Xie et al. 2020) can be interpreted as an iterative pseudo-labeling approach. Defining the model used to generate labels as "teachers" and the model trained on both pseudo-labeled and labeled dataset as "students", the two roles will be swapped in each iteration after incorporating more pseudo-labeled data into the labeled dataset (as shown in Figure 1). In this way, we can achieve our initial goal of improving model performance with the help of the valuable unlabeled dataset without human intervention or external knowledge.

3. Experiment

3.1. Datasource

The dataset we used in this paper is collected from Expertiza (Gehringer et al., 2007), a web-based peer review system that allows students to provide both numerical ratings and textual feedback to other teams' assignments. Those reviews will be used as valuable feedback for the authors to make revisions and also as an important reference for instructors to construct the final evaluation of the assignment. This reviewing work is obligatory; each student must review multiple assignments throughout the semester to earn credit, and their reviews are evaluated as well. Students have the chance to earn extra credit by labeling review comments they have received

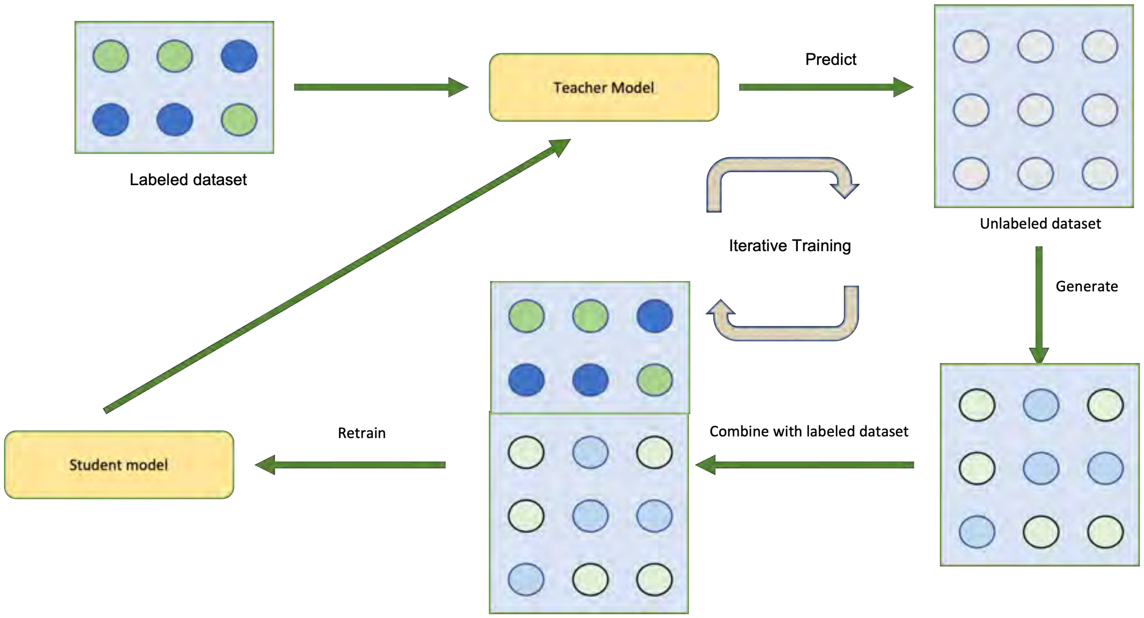

Figure 1: Self-training workflow.

from peer reviewers; these labels will be used as ground-truth labels and support the construction of text classifiers for peer-review evaluation.

For review labeling, student taggers label review comments for whether or not they exhibit characteristics such as problem statements, suggestions, and explanations. This study only uses problem-statement and suggestion labels to estimate the quality of the peer review. As mentioned previously, those ground-truth labels collected from the student taggers are not guaranteed to be of good quality. In order to extract reliable high-quality training data for the model we used the concept of inter-rater reliability (IRR) (Hallgren, 2012). Inter-rater reliability measures whether the reviewers who labeled the same comment are in agreement about whether it, e.g., detects a problem. When students used Expertiza to label comments, all members of a particular team were asked to label the same comment. We only include the labels that have been tagged by all four students in a team the same way (as detecting or not detecting a problem). As a result, 48,412 review comments labeled for problem statements and suggestions were pulled from Expertiza from the Fall 2017 to Fall 2020 semesters of a masters-level object-oriented design class. After the raw data was filtered, 3100 pieces of "high-quality" labeled data were identified. For evaluation purpose, we split these data and selected 1600 as the training set and 1500 as the validation set. The remaining 45,312 review comments without IRR=1 had their labels stripped and were used as the unlabeled dataset.

Another interesting comparison in this study looks at the effectiveness of the pseudo-labeling strategy on different sizes of the initially labeled datasets. This means instead of only feeding the entire training set (1600 labeled data) to the model, we will also randomly extract multiple subsets with different sizes as the training data. The contribution of this setting is: If we can show that the required amount of labeled data can be reduced significantly without harming the model performance, then this pseudo-labeling approach would have a great impact on peer-assessment evaluation in terms of eliminating the effort of collecting labels. For evaluation, we will focus on tracking the improvements achieved by applying our strategy with different sizes of the training labeled data, in order to determine whether our approach is effective even on an extremely small labeled dataset.

3.2. Model implementation

Comparing the performance of different language models was not the goal of this study; hence we will only select one language-classification model to train both the teacher and student models. We use the transformer-based language model known as Bidirectional Encoder Representations from Transformers (BERT), which was first introduced by Google in 2019 (Devlin et al., 2019). Transformers apply a specific self-attention mechanism, which is designed for language understanding (Vaswani et al., 2017). Self-attention emphasizes which part in an input sentence is crucial to the understanding. The transformer is an encoder-decoder-based architecture consisting of a standard feed-forward layer and a special attention layer, as shown in Figure 2 (Vaswani et al., 2017).

The traditional language model reads the input sentence in a single direction, either left to right, or right to left, which is enough for the task of next-word prediction. However, for a deep understanding of the sentence, the context is necessary. For a given word, the fact that both the previous and next token are valuable in learning the text representations, is why the BERT model can achieve such superior performance on language-understanding tasks.

The BERT model is trained in two phases, pre-training and fine-tuning. Pre-training includes two NLP tasks: Masked Language Modeling (MLM) and Next-Sentence Prediction (NSP), using 3.3 billion words from Wikipedia and BooksCorpus; note that all data is unlabeled. Then the pre-trained model is used for the downstream NLP tasks in the fine-tuning phase, like text classification. In our study, we simply used the pre-trained BERT base model, then fine-tuned the model by feeding our peer-review data to carry out the text-classification task.

3.3. Multitask learning

Multitask learning is an approach that targets multiple learning tasks at the same time instead of training multiple models for each individual task, which can potentially improve the model generalization ability by sharing lower-level features in the neural network in order to learn some common ideas among a collection of related tasks. Multitask learning is able to help achieve a better representation of the task and significantly improve training efficiency with respect to the amount of training data needed as well as the low demands on computational resources.

The inspiration of applying multitask learning in this study is the fact that problem-statement detection shares some features of the language representations with suggestion detection, and the two tasks would be able to be performed at the same time. Besides, both of these characteristics play an important role in evaluating peer reviews. We believe achieving a robust model that performs well on both tasks would be valuable. However, multitask learning is a huge research area today, and we will not dig deep into it in this study. Instead, multitask learning will only serve as an auxiliary for combining the problem statement and suggestion detection tasks, for the purpose of evaluating peer assessment more comprehensively. We will focus mainly on the

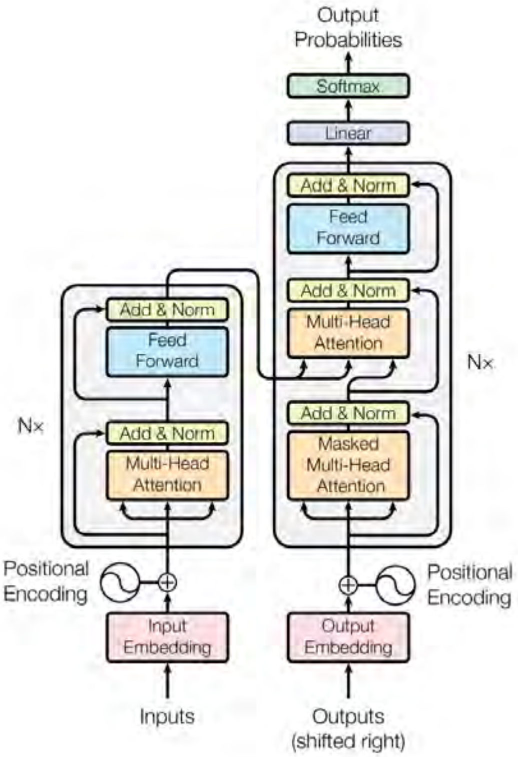

Figure 2: The Transformer - model architecture (Vaswani et al., 2017).

performance of the pseudo-labeling strategy.

As shown in Figure 3, we trained a single language model with two task heads, problem statement and suggestion detection, and then calculated the loss of the two tasks separately. We simply took an average of them as the model loss for back-propagation. In the model inference phase, we will again generate a prediction and evaluate the results for problem-detection and suggestions separately. The reason for doing that is because we propose a pseudo-labeled subset selection method that will be introduced in the following part (Section 3.4), in which the pseudo-labeled data are selected differently for the two tasks and will be incorporated into the labeled dataset.

3.4. Pseudo-labeling setting

As mentioned in Section 2.2, we initially trained a teacher model on only the labeled dataset. We aimed to evaluate the improvement achieved by our strategy with three different sizes of the labeled set. Accordingly, 400, 800, and 1600 labeled reviews were randomly selected from the initial training set. These samples were used to train the initial teacher model and later combined with pseudo-labeling data.

After the pseudo-labels are generated on the unlabeled dataset, another attractive experiment is to investigate whether we should use the entire pseudo-labeled dataset, or just a part of it, to combine with the labeled dataset and train the student model. There are three common ap-

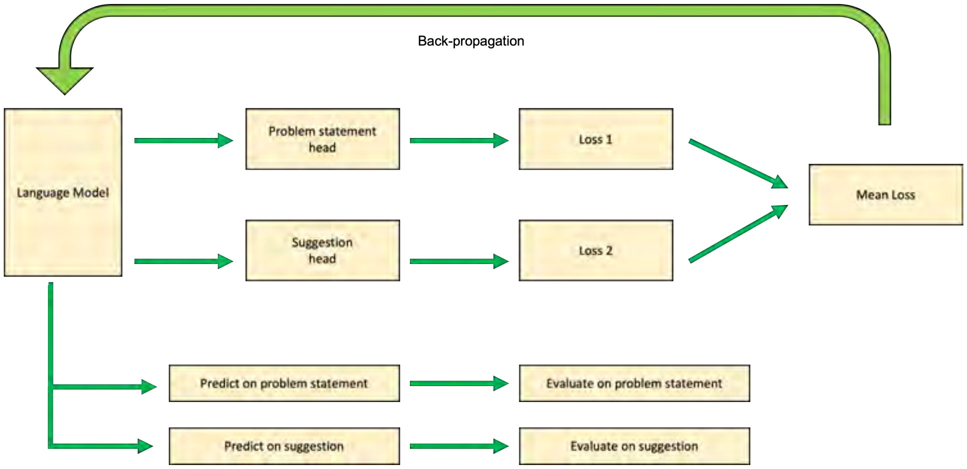

Figure 3: Multi-task Learning Flowchart.

proaches for selecting the pseudo-labeled subset. We define them as:

- **Full selection**: Combine the entire pseudo-labeled set with the labeled set to train the student model.
- **Random selection**: Randomly select a subset from the pseudo-labeled set and combine it with the labeled set; the labels of the remaining samples will be stripped, and those samples will be considered as the unlabeled dataset in the next iteration.
- **Top-k selection**: Follow almost the same steps as random selection, except for the sampling method, as shown in Figure 4. The teacher model will retain the predicting probability while generating the pseudo-labels, and then only the samples with the $k\%$ highest prediction probability (also known as highest confidence level) will be selected and incorporated into labeled dataset

In this paper, we use the top-k selection method to extract the subset of the pseudo-labeled dataset in each iteration. In order to comprehensively evaluate the validity of this probability / confidence-based sample selection method, we conducted multiple experiments with different k values (k = 10%, 20%, 40%), taking k = 100% as the baseline where no sample selection approach is applied and the entire pseudo-labeled set is directly incorporated into labeled set in a single iteration. Considering different k values results in different sample sizes selected in each iteration. For consistency, we run 10 iterations with each individual k (except for the baseline) and make sure all of the remaining pseudo-labeled data are selected in the last iteration. For the baseline (k = 100%), only one iteration is required.

As previously mentioned, pseudo-labeling is implemented as an iterative process so top-k selection will be applied repeatedly in each iteration. Once a pseudo-labeled subset has been selected, the remaining pseudo-labeled data will be stripped from the labels and used as the unlabeled set for the teacher model to make new predictions in the next iteration.

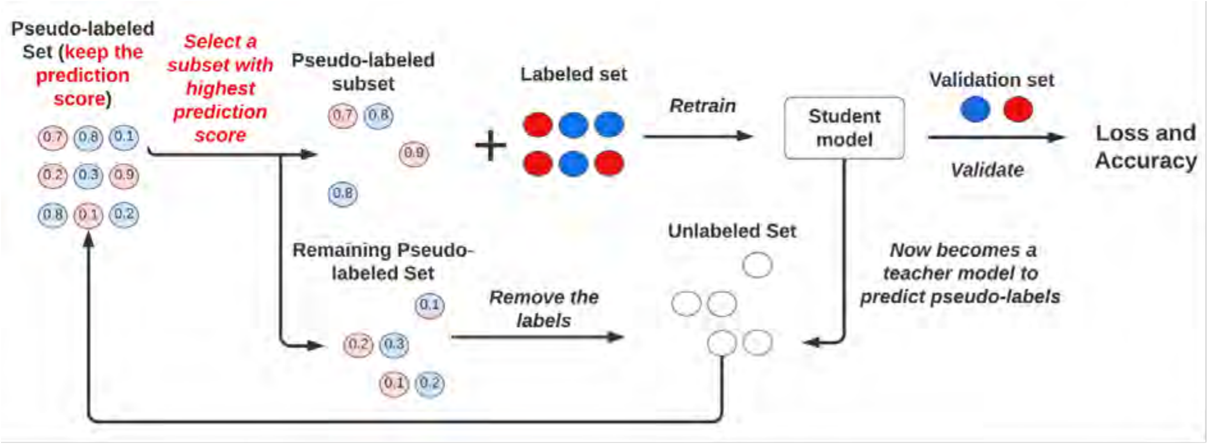

Figure 4: Top-k selection workflow.

3.5. Handling confirmation bias

Machine-learning models predict incorrect labels when they are unable to learn enough patterns from the data. In pseudo-labeling, overfitting the student model to these incorrect labels predicted by the naïve teacher model is defined as confirmation bias. This leads to a significant impairment of the pseudo-labeling strategy. Initially, the teacher model could well be affected by noise, especially with very little labeled data being trained. Although this cannot be avoided fundamentally, there are still some approaches that can help reduce the effects of confirmation bias.

3.5.1. Top-k selection

As mentioned in section 3.4, in order to eliminate the impact of incorrectly pseudo-labeled samples used during training, we aim at intelligently selecting a subset of pseudo-labels that are less noisy. The baseline approach without selecting any subset will incorporate those incorrect labelings for sure and cause training bias. Random sampling without any criteria will not be as effective at identifying incorrect labeling. Therefore, our approach attempts to filter out labels that are less likely to be accurate. During model training, the networks predict a probability that the feature is present or absent (in our case whether the comment contains or does not contain a problem statement or a suggestion). Only the data points with the highest predicted probability will be selected and included in the labeled set. The theoretical justification for this is similar to entropy regularization (Grandvalet and Bengio, 2004), which is another semi-supervised learning technique that encourages the classifier to infer confident predictions on the unlabeled dataset. For example, we would prefer to assign the unlabeled data a high probability of belonging to a particular class, rather than diffuse probabilities across different classes. However, this confidence-based approach must assume that the data are clustered according to class, which means that neighboring data points should have the same class, while the points in different classes should be widely separated.

3.5.2. Weighted loss

Another approach to handling confirmation bias is to redefine the cross-entropy loss as a weighted summation between the labeled and pseudo-labeled set. Initially, the naïve teacher model is incapable of generating reliable pseudo-labels. If we simply add the unlabeled loss to the labeled loss, especially when the size of the unlabeled dataset is much larger, the model tends to overfit on the unreliable pseudo-labeled data and consequently generate wrong predictions.
Therefore Lee et al. (2013) proposed to use weight in the loss function. The overall loss function looks like this:

$$L = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C} L\left(y_i^m, f_i^m\right) + \alpha\left(t\right)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C} L\left(y_i^{\prime m}, f_i^{\prime m}\right) \tag{5}$$

Where n and n' represents the number of samples in the labeled and unlabeled dataset respectively; C represent the number of classes, in our binary classification task C = 2; L is the binary cross entropy loss; y_i^m, f_i^m stands for the ground truth label and model prediction value on labeled dataset; $y_i^{\prime m}, f_i^{\prime m}$ stands for the pseudo label and model prediction value on unlabeled dataset.

In simple terms, the equation can be interpreted as follows:

$$Loss\,Per\,Batch = Labeled\,Loss + Weight \times Unlabeled\,Loss \tag{6}$$

In this equation, the weight (α value, $0 \leq \alpha \leq 1$) is used to control the contribution of the unlabeled loss to the total loss. This coefficient is very important to the network training. If it is set too high, the unlabeled loss is almost weighted equally to the labeled loss, and the model will be potentially fitted on incorrectly predicted labels. On the other hand, if the value is too low, the model will only learn from labeled data; thus the unlabeled data provides no benefit. Therefore, it is reasonable to set the alpha value to be small initially while the model is not learning enough patterns, then slowly increase it until we can trust the pseudo-labels. In this study, we simply initialize the alpha value to be 0.1 and increase it by 0.1 on each training iteration until the value reaches 1

4. Results and Discussion

All of the results with each experiment setting are listed in Table 1. As we can see, we first tried different initial labeled datasets with 400, 800, and 1600 for both problem statements and suggestions detection. We aimed to analyze whether the pseudo-labeling strategy is more effective on a relatively large or small labeled dataset. If we can prove that the smaller labeled dataset allows us to achieve performance similar to the large labeled dataset, then we can significantly reduce the students' tagging work and smartly assign them other tagging work to collect more labeled data. In addition, within each labeled dataset setting we also tried different k values of 10%, 20%, 40% and 100% (baseline) in the Top-k selection approach. We aimed to draw a conclusion about whether model confidence helps us avoid the influence of noise predictions.

Table 1: The improvement of accuracy and F1 score. For each size of labeled training data, the poorest improvement is shaded in red, and the best improvement is shaded in green. Note that the best improvement is always for $k < 100\%$.

	Problem Statement Accuracy			Suggestion Accuracy			Problem Statement F1			Suggestion F1		
	Init	Final	Impr	Init	Final	Impr	Init	Final	Impr	Init	Final	Impr
Training with 400 labeled data												
k=100% (Baseline)	86.97%	87.02%	0.05%	93.06%	93.48%	0.42%	84.38%	86.02%	1.64%	84.65%	85.86%	1.21%
k=10%	86.97%	89.47%	2.5%	93.06%	95.17%	2.11%	84.38%	85.80%	1.42%	84.65%	87.21%	2.56%
k=20%	86.97%	89.12%	2.15%	93.06%	94.17%	1.11%	84.38%	87.44%	3.06%	84.65%	87.32%	2.67%
k=40%	86.97%	89.06%	2.09%	93.06%	93.77%	0.71%	84.38%	87.10%	2.72%	84.65%	86.56%	2%
Training with 800 labeled data												
k=100% (Baseline)	88.93%	89.23%	0.3%	94.83%	94.75%	-0.08%	86.29%	86.85%	0.56%	87.69%	87.74%	0.05%
k=10%	88.93%	90.28%	1.35%	94.83%	95.34%	0.51%	86.29%	87.93%	1.64%	87.69%	88.66%	0.97%
k=20%	88.93%	90.15%	1.22%	94.83%	95.01%	0.18%	86.29%	87.98%	1.69%	87.69%	88.40%	0.71%
k=40%	88.93%	90.74%	1.81%	94.83%	95.53%	0.7%	86.29%	88.85%	2.56%	87.69%	90.69%	3%
Training with 1600 labeled data												
k=100% (Baseline)	89.02%	89.82%	0.8%	92.37%	93.14%	0.77%	86.63%	87.24%	0.61%	85.05%	86.37%	1.32%
k=10%	89.02%	91%	1.98%	92.37%	95.21%	2.84%	86.63%	89.09%	2.46%	85.05%	88.03%	2.98%
k=20%	89.02%	91.20%	2.18%	92.37%	94.02%	1.65%	86.63%	89.12%	2.49%	85.05%	88.78%	3.73%
k=40%	89.02%	91.46%	2.44%	92.37%	95.22%	2.85%	86.63%	89.45%	2.82%	85.05%	89.00%	3.95%

For the evaluation metrics, we tracked both accuracy and $f1$ score of the initial performance (trained only on the initial labeled set, marked as “Init”), the final performance (after applying the entire pseudo-label strategy and feeding the model exactly the same unlabeled set, marked as “Final”), and then calculated the difference to demonstrate the improvements achieved by using the approach (marked as “Impr”). This is designed to show the overall effectiveness of the pseudo-labeling strategy by comparing it with training only on the labeled dataset (not using any unlabeled dataset). The Accuracy and $F1$ values colored red and green represent the least/most improvement achieved with different k values in each labeled dataset setting. Now, let’s consider several research questions.

RQ1: Does pseudo-labeling improve the model performance?

The main contribution of this work is to demonstrate that pseudo-labeling allows us to extract valuable information and acquire new knowledge from the large unlabeled dataset without any expert supervision. We can see this from the four “Impr” columns for both Suggestions and Problem statements. All but one ($k = 100\%$ with 800 labeled sets) improved (the lack of improvement might be attributed to confirmation bias, as mentioned in the previous section and discussed below). For example, with the labeled dataset of size 400, we are able to achieve an average of 2.25% improvement in accuracy and 2.4% improvement in $f1$ score for problem statement detection

These results strongly support our hypothesis that the unlabeled set can potentially improve model performance with the semi-supervised pseudo-labeling strategy, taking advantage of plentiful unlabeled data, and relying less on labeled data, which is more difficult and expensive to collect.

RQ2: Does the top-k selection approach help improve performance?

This research question addresses the utility of our proposed top-k selection approach: Are high-confidence review comments—those rated as more likely to exhibit a certain characteristic—actually more likely to have that characteristic than low-confidence review comments? In other words, does the predicted probability of whether a review comment contains a particular characteristic (e.g., suggestion) reliably identify those comments that do exhibit that characteristic? We compare different k values to see if using lower k values to drop the low-confidence predictions will outperform the baseline approach ($k = 100\%$).

We can see from the results that while $k = 100\%$, which is the baseline, poor improvements were achieved with the entire pseudo-labeled dataset selected (the results are marked in red in Table 1). For example, training with 1600 labeled data on suggestion detection tasks without using top-k selection, we can only achieve a 0.77% improvement in accuracy and 1.32% improvement on $f1$ score. Even though we achieved some positive improvements here, considering the large unlabeled dataset we fed to the model, the improvement is unimpressive and we should be able to do much better. After applying top-k selection, we are able to achieve a maximum of 2.85% and 3.95% improvements on accuracy and f1 score, respectively. This indicates that the confidence-based top-k selection approach helps improve the model performance by avoiding the impact of incorrect predictions.

Another interesting finding here is that with different sizes of the labeled set, the best k values are different (the best improvements are marked in green in Table 1). While using the 800/1600 labeled set, $k = 40\%$ yields the best performance on all four metrics, but with the 400 labeled set, $k = 10\%$ and 20% yield the best performance on *acc* and $f1$ scores separately.

This implies that the best k value with different sizes of the labeled data is potentially related to model uncertainty. In the experiment phase, we found that 800 and 1600 labeled data seem quite enough for the model to learn the patterns, and the prediction probabilities are mostly clustered above 90%, which means that the model has little uncertainty about the inference. This leads to the result that selecting more data with similar probabilities will not harm performance but incorporate more information. However, with the 400 labeled dataset, the model has higher uncertainty and the prediction probabilities are more dispersed; this leads to more variations in the different selections of the k values. As a future experiment, adding more uncertainty to the model would potentially improve the effectiveness of the top-k selection approach.

However, our results are not quite strong enough to establish that the top-k selection approach can handle confirmation bias properly, especially with a high-confidence inference. It is natural to think that reducing the confidence of the network in its predictions might alleviate the bias problem and improve the generalization. Unfortunately, with the top-k selection alone, even though we can slightly reduce the risk of introducing much bias to the model by selecting the pseudo-labeled subset smartly, those biases from the highly confident samples that would be still propagated to the next epochs cannot be detected. Therefore, another potential approach is to combine top-k selection with some data-augmentation strategy or some modification to the model structure in order to add more uncertainty to the model for achieving a better generalization.

RQ3: Does the pseudo-labeling work better on a small labeled set or a large labeled set?

Achieving a better result with more labeled data is a very common experience in machine-learning tasks. Our approach augments a small labeled dataset with a much larger unlabeled dataset to predict characteristics using a semi-supervised approach. We have shown that it can provide performance comparable to a larger labeled dataset.
From Table 1 we can clearly see that as expected, regardless of the k value, the overall improvement on the large labeled set is actually higher than on the small labeled set. This indicates that in our work, the pseudo-labeling strategy works better by learning more patterns from more initial labeled data. Regardless of the k values, the best performance is mostly achieved with larger labeled data: for problem detection, we achieved a 91.46% best accuracy and an 89.45% best $f1$ score with 1600 labeled data, and for suggestion detection, we achieved a 95.53% best accuracy and a 90.69% best $f1$ score with 800 labeled data.

Even though none of the best accuracy/f1 scores are achieved with 400 labeled data, it is a fact that the performance is not too far away from the best results. For problem detection, it is (1.99%, 2.01%) less than the best (accuracy, f1) score; for suggestion detection, it is (0.36%m 3.37%) less than the best (accuracy, f1) score. Although we sacrifice some (accuracy, f1) scores, we achieve a 50% and a 75% reduction in the size of the labeled data which are noteworthy progress.

Therefore, we conclude that even if the pseudo-labeling approach is aimed at resolving the labeled-data insufficiency issue, a reasonable size of the initial labeled set is still required for generating more accurate inferences as well as achieving a robust model. However, from the data efficiency perspective, if a small sacrifice on the accuracy/f1 score is acceptable, we may potentially decrease the size of the labeled dataset.

5. Conclusions

This paper presents a pseudo-labeling approach for peer-assessment reviews by identifying problem statements and suggestions in the review comments, aimed at addressing the problem of insufficient labeled data. We propose a workflow combining a small amount of high-quality labeled data and a large unlabeled dataset to train a robust text classifier. For the experiment, we applied a multi-task learning strategy in order to conduct a more comprehensive peer-assessment evaluation. We evaluated the performance of this strategy with different sizes of the high-quality labeled set and applied top-k selection to handling confirmation bias (Section 3.5). The results indicate that our approach can improve the performance of the model with only a small labeled dataset by augmenting it with an unlabeled set. The main contribution of this study to the peer-review process is the revelation that not much labeled data is required to detect problem statements and suggestions in peer-assessment comments; consequently, our student taggers will not have to label so much data. With less labeled data required, student taggers can be more careful to assign correct labels.

Although we achieved good results by using the top-k selection approach as well as the weighted loss function to handle confirmation bias, there are still many potential improvements worth exploring in future investigations. Adding some uncertainty to the model training and combining with top-k selection would be a promising path to investigate. To be sure, the same success is not guaranteed on other tasks: these confidence-based selection approaches are not always applicable without the cluster assumption that similar samples in the same cluster are more likely to share a label than dissimilar samples.

The results of this study point the way to more efficiently analyzing review comments. Our pseudo-labeling experiment clearly indicates how much labeled data would be sufficient for each characteristic to train a robust text classifier. As shown in the results, too little labeled data would not be sufficient to learn enough representations for robust model inferences and would potentially be vulnerable to outliers and noise. Further research can explore better filtering approaches (like the tagger-agreement rule) for extracting small quantities of higher-quality labeled data in order to build a more reliable auto-labeling system.

REFERENCES

ARAZO, E., ORTEGO, D., ALBERT, P., O'CONNOR, N. E., AND MCGUINNESS, K. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

CALIGIURI, P. AND THOMAS, D. C. 2013. From the Editors: How to write a high-quality review. *Journal of International Business Studies 44,* 6, 547–553.

CASCANTE-BONILLA, P., TAN, F., QI, Y., AND ORDONEZ, V. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 6912–6920.

CHENG, W. AND WARREN, M. 2000. Making a difference: Using peers to assess individual students' contributions to a group project. *Teaching in Higher Education 5,* 2, 243–255.

CHO, K. 2008. Machine classification of peer comments in physics. In *Proceedings of the 1st International Conference on Educational Data Mining (EDM 2008)*, R. S. J. Baker, T. Tiffany Barnes, and J. Beck, Eds. International Educational Data Mining Society, 192–196.

DEMIRASLAN ÇEVIK, Y., HAŞLAMAN, T., AND ÇELIK, S. 2015. The effect of peer assessment on problem solving skills of prospective teachers supported by online learning activities. *Studies in Educational Evaluation 44,* March, 23–35.

DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

DOUBLE, K. S., MCGRANE, J. A., AND HOPFENBECK, T. N. 2020. The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. *Educational Psychology Review 32,* 2, 481–509.

DU, J., GRAVE, E., GUNEL, B., CHAUDHARY, V., CELEBI, O., AULI, M., STOYANOV, V., AND CONNEAU, A. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer,

D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, online, 5408–5418.

FROMM, M., FAERMAN, E., BERRENDORF, M., BHARGAVA, S., QI, R., ZHANG, Y., DENNERT, L., SELLE, S., MAO, Y., AND SEIDL, T. Argument Mining Driven Analysis of Peer-Reviews. Tech. rep.

GARCIA, R. M. C. 2010. Exploring document clustering techniques for personalized peer assessment in exploratory courses. In *Proceedings of the Workshop Computer-Supported Peer Review in Education (CSPRED-2010) held in conjunction with the Tenth International Conference on Intelligent Tutoring Systems (ITS 2010)*, I. Goldin, P. Brusilovsky, C. Schunn, K. Ashley, and I.-H. Hsiao, Eds.

GEHRINGER, E., EHRESMAN, L., GCONGER, S. G., AND WAGLE, P. 2007. Reusable Learning Objects Through Peer Review: The Expertiza Approach. *innovate Journal of On-Line education 3,* 5.

GOLDBERG, X. 2009. *Introduction to semi-supervised learning*. Vol. 6.

GRANDVALET, Y. AND BENGIO, Y. 2004. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds. Vol. 17. MIT Press, 529–536.

HALLGREN, K. A. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology 8,* 1, 23.

JIA, Q., CUI, J., XIAO, Y., LIU, C., RASHID, P., AND GEHRINGER, E. F. 2021. All-in-one: Multitask learning bert models for evaluating peer assessments. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, I.-H. Hsiao, S. Sahebi, F. Bouchet, and J.-J. Vie, Eds. International Educational Data Mining Society, 525–532.

KANG, P., KIM, D., AND CHO, S. 2016. Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing. *Expert Systems with Applications 51*, 85–106.

LEE, D.-H. ET AL. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning (WREPL) held in conjunction with the International Conference on Machine Learning (ICML)*. Vol. 3. 1–6.

LI, H., XIONG, Y., HUNTER, C. V., GUO, X., AND TYWONIW, R. 2020. Does peer assessment promote student learning? A meta-analysis. *Assessment and Evaluation in Higher Education 45,* 2, 193–211.

LIU, C., CUI, J., SHANG, R., XIAO, Y., JIA, Q., AND GEHRINGER, E. 2022. Improving problem detection in peer assessment through pseudo-labeling using semi-supervised learning. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, 391–397.

LIU, X. AND LI, L. 2014. Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment and Evaluation in Higher Education 39,* 3, 275–292.

LUNDSTROM, K. AND BAKER, W. 2009. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing 18,* 1, 30–43.

MUGNAI, D., PERNICI, F., TURCHINI, F., AND DEL BIMBO, A. 2021. Soft pseudo-labeling semi-supervised learning applied to fine-grained visual classification. In *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Springer International Publishing, Cham, 102–110.

NELSON, M. M. AND SCHUNN, C. D. 2009. The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science 37,* 4, 375–401.

NILSSON, R. 1960. A preliminary report on a boring through middle ordovician strata in western scania (sweden). *Geologiska Föreningen i Stockholm Förhandlingar 82,* 2, 218–226.

NYE, B. D., CORE, M. G., JAISWA, S., GHOSAL, A., AND AUERBACH, D. 2021. Acting engaged: Leveraging play persona archetypes for semi-supervised classification of engagement. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, I.-H. Hsiao, S. Sahebi, F. Bouchet, and J.-J. Vie, Eds. International Educational Data Mining Society, 240–251.

RAMACHANDRAN, L., GEHRINGER, E. F., AND YADAV, R. K. 2017. Automated Assessment of the Quality of Peer Reviews using Natural Language Processing Techniques. *International Journal of Artificial Intelligence in Education 27,* 3, 534–581.

SÁNDOR, A. AND VORNDRAN, A. 2009. Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. NLPIR4DL '09. Association for Computational Linguistics, USA, 36–44.

SHI, W., GONG, Y., DING, C., MA, Z., TAO, X., AND ZHENG, N. 2018. Transductive semi-supervised deep learning using min-max features. In *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer International Publishing, Cham, 311–327.

SONG, Y., HU, Z., AND GEHRINGER, E. F. 2015. Closing the circle: Use of students' responses for peer-assessment rubric improvement. In *Advances in Web-Based Learning – ICWL 2015*, F. W. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, and R. W. Lau, Eds. Springer International Publishing, Cham, 27–36.

SUEN, H. K. 2014. Peer assessment for massive open online courses (MOOCs). *International Review of Research in Open and Distance Learning 15,* 3, 312–327.

TARVAINEN, A. AND VALPOLA, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. NIPS'17. Curran Associates Inc., Red Hook, NY, USA, 1195–1204.

TOPPING, K. 1998. Peer assessment between students in colleges and universities. *Review of Educational Research 68,* 3, 249–276.

TOPPING, K. J. 2009. Peer assessment. *Theory into Practice 48,* 1, 20–27.

VAN ZUNDERT, M., SLUIJSMANS, D., AND VAN MERRIËNBOER, J. 2010. Effective peer assessment processes: Research findings and future directions. *Learning and Instruction 20,* 4, 270–279.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. U., AND POLOSUKHIN, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. NIPS'17, vol. 30. Curran Associates, Inc., Red Hook, NY, USA, 6000–6010.

XIAO, Y., ZINGLE, G., JIA, Q., AKBAR, S., SONG, Y., DONG, M., QI, L., AND GEHRINGER, E. 2020. Problem detection in peer assessments between subjects by effective transfer learning and active learning. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, Eds. International Educational Data Mining Society, 516–523.

XIAO, Y., ZINGLE, G., JIA, Q., SHAH, H. R., ZHANG, Y., LI, T., KAROVALIYA, M., ZHAO, W., SONG, Y., JI, J., ET AL. 2020. Detecting problem statements in peer assessments. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, A. N. Rafferty,

J. Whitehill, C. Romero, and V. Cavalli-Sforza, Eds. International Educational Data Mining Society, 704–709.

XIE, Q., LUONG, M. T., HOVY, E., AND LE, Q. V. 2020. Self-training with noisy student improves imagenet classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10684–10695.

XIONG, W. AND LITMAN, D. 2010. Identifying problem localization in peer-review feedback. In *Intelligent Tutoring Systems*, V. Aleven, J. Kay, and J. Mostow, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 429–431.

XIONG, W. AND LITMAN, D. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Association for Computational Linguistics, Portland, Oregon, USA, 502–507.

XIONG, W., LITMAN, D., AND SCHUNN, C. 2010. Assessing reviewers' performance based on mining problem localization in peer-review data. In *Proceedings of the 3rd International Conference on Educational Data Mining (EDM 2010)*, R. S. Baker, A. Merceron, and P. I. Pavlik, Eds. International Educational Data Mining Society, 211–220.

ZHOU, Z. H. 2018. A brief introduction to weakly supervised learning. *National Science Review 5,* 1, 44–53.

ZHOU, Z.-H. AND LI, M. 2005. Semi-supervised regression with co-training. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 908–913.

ZINGLE, G., RADHAKRISHNAN, B., XIAO, Y., GEHRINGER, E., XIAO, Z., PRAMUDIANTO, F., KHURANA, G., AND ARNAV, A. 2019. Detecting suggestions in peer assessments. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, C. F. Lynch, A. Merceron, M. Desmarais, and R. Nkambou, Eds. International Educational Data Mining Society, 474–479.