Assessing the Quality of Student-Generated Content at Scale: A Comparative Analysis of Peer-Review Models

Ali Darvishi, Hassan Khosravi, Afshin Rahimi, Shazia Sadiq, and Dragan Gašević

***Abstract*—Engaging students in creating learning resources has demonstrated pedagogical benefits. However, to effectively utilize a repository of student-generated content (SGC), a selection process is needed to separate high- from low-quality resources as some of the resources created by students can be ineffective, inappropriate, or incorrect. A common and scalable approach is to use a peer-review process where students are asked to assess the quality of resources authored by their peers. Given that judgments of students, as experts-in-training, cannot wholly be relied upon, a redundancy-based method is widely employed where the same assessment task is given to multiple students. However, this approach introduces a new challenge, referred to as the consensus problem: How can we assign a final quality to a resource given ratings by multiple students? To address this challenge, we investigate the predictive performance of 18 inference models across five well-established categories of consensus approaches for inferring the quality of SGC at scale. The analysis is based on the engagement of 2141 undergraduate students across five courses in creating 12 803 resources and 77 297 peer reviews. Results indicate that the quality of reviews is quite diverse, and students tend to overrate. Consequently, simple statistics such as mean and median fail to identify poor-quality resources. Findings further suggest that incorporating advanced probabilistic and text analysis methods to infer the reviewers' reliability and reviews' quality improves performance; however, there is still an evident need for instructor oversight and training of students to write compelling and reliable reviews.**

***Index Terms*—Consensus approaches, crowdsourcing in education, learnersourcing, learning analytics, peer review.**

I. Introduction

LEARNERSOURCING refers to a pedagogically supported form of crowdsourcing that mobilizes the learner community as experts-in-training to contribute novel content for future learners while being engaged in meaningful learning experiences themselves [1], [2]. This emerging concept has been inspired by the success of crowdsourcing as an effective problem-solving paradigm that leverages the crowd for primarily completing a task. The concept of engaging learners as contributors to novel content has strong roots in the learning sciences and is aligned with established and contemporary learner-centered approaches [3]. One important approach in which learnersourcing has been used is to harness the collective creative power of students to develop repositories of learning resources [4], [5], [6].

The development of repositories of student-generated content (SGC) can have various benefits: It can be used by 1) students for studying [4], 2) instructors for creating assessments [7], or for engaging students in higher-order learning tasks [8], as well as 3) adaptive educational systems, which need large repositories of learning resources, for recommending personalized instructions [5], [9]. However, to effectively use SGC repositories, high- and low-quality resources should be identified through moderation. While strong evidence from previous work suggests that a large portion of the SGC is of high quality and meets rigorous judgmental and statistical criteria [10], [11], it also suggests that students commonly create resources that are ineffective, inappropriate, or incorrect [7], [12], [13].

So how can we separate high-quality from low-quality resources in a large SGC repository? One approach is to engage instructors as experts in evaluating the quality of the resources; however, the instructor-led quality evaluation is not scalable and can be expensive due to the potentially large size of the repositories. An alternative scalable approach is to use a peer-review process where students are asked to assess the quality of resources authored by their peers. Engaging students in providing instead of receiving feedback can be beneficial for learning [14], [15] and has the capacity to help students develop evaluative judgment, which has been recognized as an important aspect of the learning process [16]. Although some prior work has reported on learners' ability to evaluate resources effectively [7], [11], [17], [18], the judgments of students as experts-in-training cannot wholly be trusted. A common solution is to rely on the wisdom of the crowd rather than one person by employing a redundancy-based strategy and assigning the same task to multiple users. This solution has been also utilized in other types of assessments such as peer reviewing of academic manuscripts [19] or grant proposals [20] and more widely by the

Manuscript received 22 September 2021; revised 24 July 2022; accepted 3 December 2022. Date of publication 14 December 2022; date of current version 17 February 2023. This work was supported by the Australian Government through the Australian Research Council's Industrial Transformation Training Centre for Information Resilience (CIRES) under Grant IC200100022. *(Corresponding author: Ali Darvishi.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The University of Queensland under Application No. 2018000125, and performed in line with the Human Research.

Ali Darvishi, Hassan Khosravi, Afshin Rahimi, and Shazia Sadiq are with the School of Information Technology and Electrical Engineering, The University of Queensland, QLD 4072, Australia (e-mail: a.darvishi@uq.edu.au; h.khosravi@uq.edu.au; a.rahimi@uq.edu.au; shazia@itee.uq.edu.au).

Dragan Gašević is with the Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia (e-mail: dragan.gasevic@monash.edu).

Digital Object Identifier 10.1109/TLT.2022.3229022

crowdsourcing community [21]. However, it raises a new problem commonly referred to as the consensus problem: In the absence of ground truth, how can we optimally integrate the decisions made by multiple people toward an accurate final decision?

In response to this question, we investigate the predictive performance of five categories of consensus approaches for evaluating the quality of SGC based on a peer-review process. The first category relies on summary statistics such as mean or median. This category is chosen as summary statistics are commonly used consensus approaches within many crowdsourcing and educational peer-review systems (e.g., [22], [23], [24], [25]). The second category relies on historical performance or self-evaluation data to approximate student competence and reliability. This category is chosen as student competence models are commonly used in adaptive educational systems (e.g., [9]) for approximating students' abilities. The third category incorporates probabilistic consensus approaches that infer the reliability of students based on their past reviews. This category is chosen as it is commonly and successfully used in crowdsourcing systems (e.g., [26], [27], [28]) to approximate the reliability of crowd workers. The fourth category incorporates text analysis methods that infer the reliability of a review based on the provided comment. This category is chosen as it is commonly and successfully used in the context of identifying reliable reviews (e.g., [29], [30], [31], [32]). The fifth category combines approaches from the previous four discussed categories. It is chosen due to the success of ensemble consensus approaches (e.g., [33], [34]). Our investigation is guided by the following research questions.

RQ1. How well do the commonly used summary statistics infer the quality of SGC in the peer-review process?

RQ2. To what extent is student judgment of content quality associated with their learning competence history and self-assessment of confidence?

RQ3. How does inferring the reliability of reviewers by probabilistic models impact the SGC quality inference?

RQ4. How does inferring the reliability of a review by text analysis models impact the SGC quality inference?

RQ5. Do combinations of the above models improve the performance of the SGC quality inference?

These research questions are answered by comparing the predictive performance of 18 consensus approaches across the five presented categories. The performance of the models is evaluated based on the engagement of 2141 students across five undergraduate courses using empirical data collected from the adoption of a learnersourced system. The rest of the article is organized as follows. The following section presents related work on learnersourcing and consensus approaches. Section III presents a suite of consensus inference models to infer resource quality. Section IV describes the tool, data, and metrics used to answer the above RQs. Then, Section V compares and contrasts the performance of a suite of consensus approaches against four metrics on peer review of SGC quality. Section VI discusses the implications and potential benefits and shortcomings of integrating the presented algorithms into an educational system. We also suggest directions to pursue in future work to overcome current limitations. Finally, Section VII concludes this article.

II. Literature Review

We explored the literature in two different realms: 1) The next section covers prior work related to learnersourcing, and 2) then various consensus inference methods are presented.

A. Learnersourcing

Learners benefit from learnersourcing both when they engage with resources and when they create resources themselves. This is supported by learner-centered theory [3] and generation effect [35]. The advantages of self-generated information on learning and memory have been acknowledged over reading the information provided by others [36], [37]. A growing number of systems enable students to create various forms of content, such as knowledge components [38], multiple-choice questions [4], [39], personalized hints [40], summaries of steps in how-to videos [41], explanations for peer instruction [42], solutions to open-ended questions [43], and explanations for programming misconceptions [44]. One of the listed challenges of learnersourcing systems is how they can control the quality of the created content at scale [45]. Peer-review systems have generally addressed this challenge by a redundancy-based strategy that distributes the quality evaluation task to multiple peers.

The role of high-quality feedback in learner outcomes is well attested in educational research [46], [47]. However, it is hard to scale it to the large number of items that need to be assessed in many learning platforms. Peer assessment not only scales well but also promotes a higher level of learning compared to one-way instructor assessment [48], [49], [50]. Peer evaluation activities range from involving more experienced learners to help novices with hints and reviews to pairing students to assess each other's activities or flagging an activity to be further assessed by instructors. These peer evaluation methods are implemented in a number of learning platforms such as Mechanical TA [22], Dear Beta and Dear Gamma [40], Aropä [23], PeerScholar [24], CrowdGrader [51], edX [52], and Peergrade [25]. Measuring the quality of peer evaluations is a major challenge for their integration into regular educational programs (e.g., for formative and summative marking), which is the focus of this work. In the following section, we briefly discuss the commonly applied methods for making a final decision from multiple peer reviews.

B. Consensus Approaches

In the crowdsourcing literature, the problem of optimal integration of crowdsourced decisions in the absence of a ground truth toward making an accurate final decision has been studied under the general terms of truth inference or consensus approaches [21], [53]. Here, we focus on the following five groups of models for estimating the quality of SGC:

1) summary statistics that only rely on user ratings;

2) learner competence models that estimate students' ability based on historical student performance or self-evaluation;
3) probabilistic models that estimate moderators' reliability or predict their behavior to adjust their contributions toward consensus;
4) text analysis models that estimate moderation's reliability based on the comments provided by peer reviewers;
5) models that combine approaches from the previous four discussed categories.

1) Summary Statistics: These models use summary statistics of the decisions given by a crowd on each item to infer a final rating. These are explainable, and users can easily understand the outcomes. Summary statistics methods like mean aggregation are commonly employed in peer evaluation systems such as Mechanical TA [22], Aropä [23], PeerScholar [24], and Peergrade [25] to integrate student decisions on their peer's work. However, these methods are quite fragile against skewed data and users with diverse abilities or interests [54]. A number of studies have attempted to extend the baseline methods by using weighted summaries and optimization techniques [55], [56]. For example, Weir et al. [41] used a majority voting approach to determining the final subgoal labels from learner decisions on instructional videos in Crowdy. While Weir et al. [41] reported that most of the learners' evaluations were comparable to experts, they also employed a multistage approach for proofreading to tackle low-quality (spam) annotations. Similarly, Williams et al. [57] used a weighted average approach to evaluate the helpfulness of provided explanations in AXIS. They also claimed that the quality of explanations in their system was comparable to that of a skilled academic. In this study, we use four summary statistics methods (majority voting, mean, median, and debiased mean) as baselines and compare their performance with learner competence, probabilistic, and text analysis models.

2) Learner Competence Models: A notable limitation of summary statistics is that each user's contribution has the same impact on the final result, whereas the quality of ratings and user reliability may vary substantially across a cohort [58], [59]. The problem of unfair representation is addressed by incorporating a competence-weighted approach in crowdsourcing systems, where more skilled workers would receive more weight than others in the crowd [60]. Tao et al. [61] emphasized that crowdsourced labeling systems should utilize a weighted majority vote method to aggregate the noisy labels so that higher competent annotators are given greater weight in the final decision. In educational systems, students' competency is commonly utilized to adapt instructions [5], [9], [62] or build learner models in the system [63], [64]. Abdi et al. [58] used auxiliary data from student performance in an unsupervized learnersourcing consensus approach to improve the accuracy of determining the quality of learning resources. In this study, we use three methods for estimating a learner's competency and then use it in a way that more competent and engaged students have a larger contribution to the final decision.

3) Probabilistic Inference Models: Estimating the competence level of users in crowdsourcing systems is a challenging task due to the absence of ground truth labels, and the anonymity of annotators [61]. Therefore, user reliability is commonly inferred using data-driven latent models in the absence of ground truth [26]. One of the well-adopted probabilistic models to estimate the quality of response is expectation–maximization (EM), which is used as a weighted aggregation method in consensus approaches [27]. Whitehill et al. [17] developed a probabilistic model using EM to estimate the quality of learning resources by aggregating learners' subjective ratings. However, users' inherent anonymity in large-scale networks such as crowdsourcing platforms and social media raises malicious behavior such as spamming and providing false or misleading information. This misbehavior is considered challenging to detect by EM [54]. Therefore, trust evaluation in large-scale networks becomes vital to tackle this challenge [65]. In this regard, trust propagation approaches are probabilistic models that try to disclose spammers and untrustworthy users in social networks [65]. A review graph model proposed by Wang et al. [66] identifies untrustworthy online store reviewers using an iterative approach. Guha et al. [67] introduced a trust propagation framework that includes distrust. Besides, several probabilistic models consider the consensus inference as an information recommendation problem and use collaborative filtering recommendation methods like matrix factorization (MF), item-based collaborative filtering, and tensor factorization to reach consensus [28], [68], [69], [70]. MF has been shown as a resolution to deal with the sparsity of user ratings in product reviews or movie rating applications [68], [70] and learner moderation on resources that are particularly sparse. This study evaluates three probabilistic methods—EM, trust propagation, and MF, to estimate student moderators' reliability in the peer-review process.

4) Text Analysis Models: While research on automatic estimation of peer feedback quality in educational systems is scarce [71], there is a large body of work in natural language processing (NLP) on estimating the quality or helpfulness of product reviews, which can be adapted for peer feedback quality estimation. Automatically estimating peer feedback reviews can be formulated as either a text categorization (or regression, depending on the desired output) when feedback is considered in isolation, or feedback-resource pair categorization when the relationship between feedback and its resource is taken into account. Features such as the number of tokens, sentences and question/exclamation marks indicating effort, count of positive and negative words, specialized features such as content localization phrases for long essay feedback (e.g., in page 6), and modal verbs (e.g., must, could) are used for predicting the helpfulness of product reviews [29] and peer feedback quality [30] in isolation. Xiong and Litman [30] report that feedback length correlates the most with feedback quality among these features. Additionally, Duret et al. [72] report that students who engage with longer comments have better improvement in learning outcomes compared to others. In this work, we also use feedback text length to assess the quality of peer evaluations. Features from review–product pairs, such as the relatedness of reviews to the corresponding product, have been

used for helpfulness prediction. Zhang and Varadarajan [31] reported a high dependence between the perceived efficacy of a product review and its linguistic style. Recent works have used supervised end-to-end neural architectures to identify the helpfulness of product–review pair [73]. Devlin et al. [74] introduced BERT, a neural language model pretrained on a large language corpus, to encode sentences such that pairs of related sentences (e.g., with similar meaning) are close to each other in the embedding space. Xu et al. [75] use BERT in measuring online product reviews' quality in a supervised setting. However, given that the amount of annotated data in peer-review assessments is very limited, an unsupervised neural architecture—Sentence-BERT (SBERT) [76], is employed in this study to identify the relatedness of feedback to the learner-sourced content. SBERT measures comment-resource relatedness with no supervision, which is very important given the variation, scale of peer-provided feedback, and the lack of annotation. This work evaluates the usefulness of four text analysis features, including sentiment alignment, length, similarity, and relatedness.

5) Combined Models: Combining multiple models, also referred to as ensemble modeling, is used to improve the prediction performance in machine learning and crowdsourcing literature [33]. There is overwhelming evidence that the use of ensemble methods also improves predictive performance on imbalanced datasets [34]. For example, in a system for topic labeling of multimedia posts, Chang et al. [77] show that an ensemble model that first learned the reliability of annotators from crowdsourced judgments outperformed a naive method that aggregates labels from annotators. Zhang et al. [78] also proposed ensemble solutions based on majority voting and maximum likelihood estimation to predict unlabeled data by aggregating multiple base classifiers, which showed to outperform a set of advanced algorithms. Although the outperformance of multimodal models often comes at the expense of increased computational requirements and reduced explainability [79], their success in machine learning and crowdsourcing problems motivates us to investigate the performance of various combinations of the features mentioned above, especially to find a multimodal model that considers the reliability of both reviewers and reviews.

III. Methods

Here, we first present a formal definition and notation for the problem under investigation. We then present 18 representative models from the 5 categories of consensus approaches. Table I provides a summary of the notations used in this article.

A. Problem Definition and Notation

Let $U_N = \{u_1 \dots u_N\}$ denote a set of users enrolled in a course in an educational system, where u_i refers to an arbitrary user. Let $Q_M = \{q_1 \dots q_M\}$ denote a set of learning resources, where q_j refers to an arbitrary resource. A resource q_j either holds a moderated status, which means its quality $\hat{r}_j$ has been inferred and exceeds a threshold of Γ, or it holds a nonmoderated status, which means its quality $\hat{r}_j$ is unknown. Peer reviewing a nonmoderated resource by users includes providing a decision rating about the quality, accompanied by a comment rationalizing their decision and a confidence rating to assess their confidence in their decision. Let $D_{N\times M}$ capture users' decision ratings on evaluation of learning resources where $1 \leq d_{ij} \leq 5$ shows the decision rating given by user u_i to resource q_j. Let $C_{N\times M}$ denote comments provided by users to accompany decision ratings, where c_{ij} stores the comment of user u_i on resource q_j, and let $\Phi_{N\times M}$ denote confidences that are provided to accompany decision ratings, where $1 \leq \phi_{ij} \leq 5$ is the confidence level of user u_i on their rating for resource q_j.

TABLE I
Notation Used in the Problem Definition and the Presented Approaches

Input Parameters	
U_N	A set of users $\{u_1 \dots u_N\}$ who are enrolled in a course, where u_i is an arbitrary student.
Q_M	A repository of learning resources $\{q_1 \dots q_M\}$ available within the system, where q_j is an arbitrary resource.
$D_{N\times M}$	A two dimensional array in which $1 \leq d_{ij} \leq 5$ shows the decision rating given by user u_i to resource q_j.
$C_{N\times M}$	A two dimensional array in which c_{ij} denote the comment provided by user u_i on resource q_j.
$\Phi_{N\times M}$	A two dimensional array in which $1 \leq \phi_{ij} \leq 5$ shows user u_i self-assessment of confidence level in rating resource q_j.
Γ	A threshold of quality for moderated resources.
Inference Models Parameters	
B_N	A set of users' bias $\{b_1 \dots b_N\}$ in which b_i shows the bias of user u_i in rating the quality of resources.
$\bar{d}_i$	The average decision rating of user u_i.
$\bar{d}$	The average decision rating across all users.
$LC_{N\times M}$	A two dimensional array in which lc_{ij} denote the length of the comment provided by user u_i on resource q_j.
$F^A_{N\times M}$	A function where f^A_{ij} approximates the alignment between the rating and comment provided by u_i on q_j.
$F^R_{N\times M}$	A function where f^R_{ij} determines the quality of the rating provided by u_i for q_j.
$S_{N\times N}$	A two dimensional array in which $-1 \leq s_{ik} \leq +1$ shows the similarity value between user u_i and user u_k.
Λ_N	A set of users' reliability score $\{\lambda_1 \dots \lambda_N\}$ in which λ_i infers the reliability score of user u_i .
ρ	The initial value of the moderation score for all users.
Output	
$\hat{R}_M$	A set of M ratings $\{\hat{r_1} \dots \hat{r_M}\}$ where each rating $1 \leq \hat{r}_j \leq 5$ shows the quality of resource q_j.

B. Consensus Problem Definition

Given a nonmoderated resource q_j and evaluations $d_{1j}, \dots, d_{kj}$, infer the quality of q_j denoted as $\hat{r}_j$.

C. Summary Statistics

The consensus approaches presented in this subsection only use aggregate statistics over the numerical ratings provided by students, which makes them fast (can be used in real-time) and easily explainable.

1) Majority Vote: A common consensus approach is to use a majority vote or mode, which takes the rating given by the majority as the outcome.[1] Favorable characteristics of this approach are that it can be used on both categorical and

[1] Ties are broken via a random assignment of one of the ratings competing for the majority as the outcome.

numerical values, is easy to explain, and is generally viewed as a fair approach.

2) Mean: A simple consensus approach based on summary statistics is to use the mean of all ratings: $\hat{r}_j = \frac{\sum_{i=1}^{k} d_{ij}}{k}$. In mean aggregation, the same contribution weight is given to all moderators in inferring the final rating.

3) Median: Another rating aggregation method we used is $\hat{r}_j = \text{Median}(u_1, \dots u_k)$. Median often lies between mean and mode in skewed normal distributions. Similar to the mean aggregation, it assumes an equal weight for all ratings and also ignores user bias.

4) Debiased Mean: A considerable number of students consistently underestimate or overestimate the quality of resources. Incorporation of under- and overrating into the consensus approach can reduce its impact. We introduce the notation of B_N, where b_i shows the rating bias of user u_i. Introducing a bias parameter has been demonstrated to be an effective way of handling user bias in several domains, such as recommender systems and crowd consensus approaches [80]. In the current study, we first computed $\bar{d}_i$ as the average decision rating of user u_i. We then computed $\bar{d} = \frac{\sum_{i=1}^{N} \bar{d}_i}{N}$ as the average decision rating across all users. The bias term for user u_i was computed as $b_i = \bar{d}_i - \bar{d}$. Positive values of b_i indicate that, on average, u_i ratings compared to their cohort were higher. Conversely, negative values of b_i show that u_i underrated resources compared to their cohort. Rating bias was prevalent in our data and negatively impacted aggregation, so needs to be accounted for. To adjust for the rating bias, the quality of resource q_j can be inferred as $\hat{r}_j = \frac{\sum_{i=1}^{k} (d_{ij} - b_i)}{k}$, where k is the number of ratings on resource q_j.

D. Learner Competence

The learner competence models aim to use the available data on student moderators to approximate competence, which in turn is used to infer their reliability. Using (1), learner competence models utilize the following general formula:

$$\hat{r}_j = \frac{\sum_{i=1}^{k} \lambda_i \times d_{ij}}{\sum_{i=1}^{k} \lambda_i} \tag{1}$$

where λ_i is representing the competency estimates of the learner i. The number of resources a student has correctly answered before rating resource q_j is used for approximating λ_i as their competence, which can be collected by analyzing their past contributions at the course level or specific topics associated with the target resource q_j. We also consider students' Elo score, a self-correcting rating from the learner model utilized in the educational system that reflects students' knowledge states and learning resources difficulty over time [64].

1) Learner Confidence: An alternative approach of approximating a student's reliability in evaluating a resource is to set λ_i to the self-provided confidence level ϕ_{ij} of the student so that more confident ratings contribute more to the final rating.

E. Probabilistic

Many methods have been introduced for computing reliability of users [21]. Here, we adopt three popular probabilistic models for inferring moderator reliability.

1) Expectation-Maximization: The problems of inferring the reliability of users Λ_N and the quality of resources $\hat{R}_M$ can be seen as solving a "chicken-and-egg" problem where inferring one set of parameters depend on the other. If the true reliability of students Λ_N were known, then an optimal weighting of their decisions could be used to estimate $\hat{R}_M$. Similarly, if the true quality of resources $\hat{R}_M$ were known, then the reliability of each student Λ_N could be estimated. In the absence of ground truth for either, the following procedure inspired by the well-adopted EM technique is used:

1) set the reliability of all students to an initial value of ρ;
2) infer $\hat{r}_j$ for a resource q_j based on current values of learner reliability scores $\lambda_1, \dots \lambda_k$ and, ratings $d_1, \dots d_k$ on q_j;
3) update $\lambda_1, \dots \lambda_k$.

In this method, the current ratings of the users and their given decisions are utilized for computing the quality of the resources and reliability of the users using (2) as follows:

$$\hat{r}_j = \frac{\sum_{i=1}^{k} \lambda_i \times d_{ij}}{\sum_{i=1}^{k} \lambda_i}, \quad \lambda_i = \lambda_i + f_{ij}^{R}(d_{ij}, \hat{r}_j) \tag{2}$$

where $f_{ij}^{R}(d_{ij}, \hat{r}_j) = \frac{2\delta e^{-(d_{ij}-\hat{r}_j)^2/(2\sigma^2)} - \delta}{2\sigma\sqrt{2\pi}}$ determines the "goodness" of d_{ij} based on $\hat{r}_j$ as the height of a Gaussian function at value $(d_{ij} - \hat{r}_j)$ with center 0, standard deviation $\sigma = .7$ and peak $\delta = 100$.

2) Graph-Based Trust Propagation: In this approach, Wang et al.'s [66] review graph model and Guha et al.'s [67] trust propagation framework are merged to consider agreement and disagreement to compute the similarity between users and then estimate and propagate users' reliability [59]. As shown in Fig. 1(a), a graph is considered that consists of four kinds of nodes—students, decision ratings, resources, and instructors. This graph-based trust propagation model has three main stages: 1) decision-making, 2) updating scores, and 3) reliability propagation. The first two stages are very similar to the EM method. An initial score (e.g., ρ) is devoted to all students. This set of scores is transformed into an initial set of users' reliability Λ_N. In the decision-making stage, given a nonmoderated resource q_j and a set of students' decision rating $(d_{1j}, \dots, d_{kj})$, the system estimates the quality when enough reliable and trustworthy moderators have evaluated the given resource. Then, in the updating score stage, the inferred quality of q_j (i.e., $\hat{r}_j$) is used to calculate the moderators' gained score. Finally, in the propagation stage, all other users connected to this set of users (i.e., $(u_1, \dots, u_k)$) would also receive an updated score from the most reliable and trustworthy moderator. In this scenario, users' reliability would be updated by the amount of φ_i, which depends on the quality of their own work and also similarities s_{ik} to their peers who are directly connected to them as a result of their collaboration in the previous moderations, as shown in Fig. 1(b).

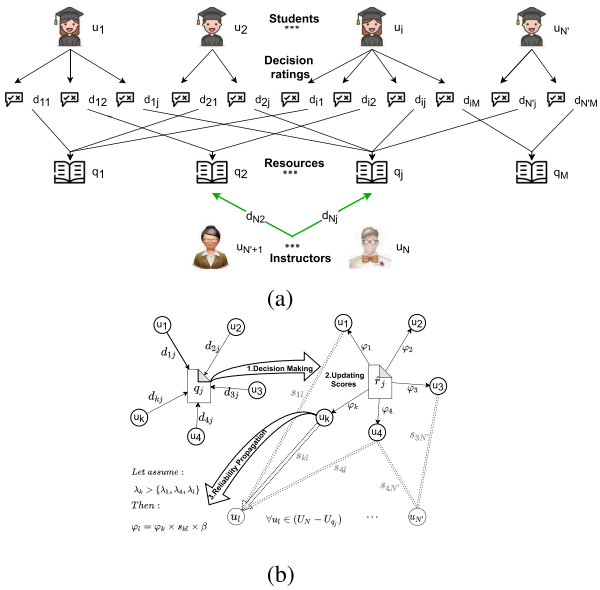

Fig. 1. Graph-based trust propagation. (a) Moderation graph with four kinds of nodes—students, decision ratings, resources, and instructors, and (b) main steps in the propagation network model.

$$D_{N\times M}=\begin{bmatrix} & q_1 & \cdots & q_j & \cdots & q_M & \\ & d_{11} & & d_{1j} & & d_{1M} & u_1 \\ & & \ddots & & \cdots & & \vdots \\ & d_{i1} & & d_{ij} & & d_{iM} & u_i \\ & \vdots & & & \ddots & & \vdots \\ & d_{(N-1)1} & & d_{(N-1)j} & & d_{(N-1)M} & u_{N-1} \\ & d_{N1} & & d_{Nj} & & d_{NM} & u_N \end{bmatrix}$$

$$\begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 & \\ 5 & 3 & & 1 & & u_1 \\ 4 & & & 1 & 3 & u_2 \\ 1 & 1 & 2 & & & u_3 \\ 1 & & 4 & 5 & 1 & u_4 \\ & 1 & & 4 & ? & u_N \end{bmatrix} \rightarrow \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 & \\ 4.97 & 3.01 & 3.36 & 1.01 & 3.49 & u_1 \\ 3.98 & 1.59 & 2.32 & 1.01 & 3.00 & u_2 \\ 1.06 & 0.92 & 2.05 & 1.71 & 0.49 & u_3 \\ 0.99 & 2.31 & 3.98 & 4.99 & 1.03 & u_4 \\ 1.90 & 1.08 & 2.77 & 3.98 & \mathbf{2.08} & u_N \end{bmatrix}$$

Fig. 2. Overview of the rating matrix passed to MF using student and instructor ratings with an example.

3) Matrix Factorization (MF): Available user moderation per resource is a sparse matrix since each student evaluates a few resources. Here, to infer students' and instructors' estimated ratings on the quality of all resources, a semisupervized MF approach [70] is used to induce latent feature vectors. In this approach, a rating matrix, as shown in Fig. 2, is constructed so that the first $N-1$ rows consist of the decision ratings of students on the quality of the M resources and the last row (N) is for the instructor ratings. For each resource, MF would be applied to estimate how instructors would evaluate the quality of that resource. The estimated quality is recorded to be compared with the actual instructors' rating in the test (spot-checked) set for evaluation and comparison. Then, for the following resources, the previous elements of the last row would be populated by the available instructor's decision.

F. Text Analysis

The previous numerical rating and learner competency models take into account the similarity of the students' numeric rating with those of their peers, but they do not take into account how much effort was applied by a user in the evaluation of a resource. When students moderate a resource, in addition to the numerical rating and confidence, they provide textual feedback. We describe three methods to incorporate comments into the consensus approach: 1) comment length, 2) sentiment-rating alignment, and 3) relatedness/similarity of the comment to the resource.

1) Length: The amount of effort learners spent on moderation can be measured by the length of their comments. The notation of $\mathrm{LC}_{N\times M}$ is introduced, where lc_{ij} shows the length of comments (i.e., number of words) provided by user u_i on resource q_j. The final rating $\hat{r}_j$ is inferred using (1), where λ_i is set to lc_{ij} which approximates the "effort" of u_i in answering q_j based on the length of the comment. Informally, this model rewards students that have provided a longer explanation for their rating.

2) Sentiment-alignment: In many cases, students might provide a comment that criticizes the resource (negative) in the comment (e.g., "This question is confusing, you are mixing concept X with Y.") but rate the resource high. In other cases, the students provide a positive unhelpful comment like "very good" or "great" but provide a low rating. The lack of alignment between a comment's sentiment and rating is often indicative of a lack of effort or misunderstanding of the rating scheme. The sentiment-alignment model measures a provided comment's sentiment and scores its alignment with the given rating to address this issue. In this method, the final rating is inferred by (1) by setting λ_i to f^A_{ij}, where $F^A_{N\times M}$ is a function approximating the alignment of the rating d_{ij} and the comment c_{ij} a user u_i has provided for a resource q_j. A sentiment analysis tool, Jockers–Rinker sentiment lexicon, classifies words in peer feedback into positive, negative, and neutral sentiments and then computes sentiment score [81], [82]. This tool assigns polarity to words in strings with valence shifters. For example, it would recognize this sample comment "This question is not useful for this course" as a negative rather than indicating the word "useful" as positive.

3) Relatedness: Analysis of feedback text in isolation might be indicative of student effort, but it does not necessarily show how feedback relates to the corresponding resource. To analyze this relation, features from the feedback–resource pair should be used. This usually involves measuring the number of exact overlapping words between the pair; however, the words used in a related comment might not exactly be the words used in the corresponding resource. Linguistic variation, such as using plurals, synonyms, or hyponyms, can result in a word mismatch. A comment that relates and mentions aspects of the resource under review is more insightful and indicative of critical thinking compared to a general comment such as "very good." To find the relatedness of a comment c_{ij}

to a resource q_j, both c_{ij} and q_j are first encoded in a semantic vector space, and then, their cosine similarity is measured in that space

$$\vec{c}_{ij} = \text{Encoder}(c_{ij}) \quad \& \quad \vec{q}_j = \text{Encoder}(q_j)$$
$$\text{Relatedness}(c_{ij}, q_j) = \cos(\vec{c}_{ij}, \vec{q}_j). \tag{3}$$

To capture semantic relatedness rather than only relying on exact lexical matches, SBERT [76] is used as the encoder function in (3). The cosine similarity score [83], [84], [85] between the two representations measures relatedness and ranges in $[-1, 1]$. SBERT is based on BERT [74] and is pretrained to encode text into semantically meaningful representations. Given that these models are pretrained on large amounts of English text, they can be used with little or no supervision in many NLP tasks with state-of-the-art performance. In addition, GLEU (Google BiLingual Evaluation Understudy) is used as a measure of similarity using n-grams between the provided comment and the resource under moderation [86]. After computing the relatedness and similarity of comment–resource pair, (1) is used to measure the final rating $\hat{r}_j$ by setting λ_i to normalized $\cos(\vec{c}_{ij}, \vec{q}_j)$, the relatedness of comment c_{ij} on resource q_j, and similarity score based on the GLEU.

G. Combined Models

The consensus approaches described above utilize various information collected from student moderators to infer the quality of a resource. Combining features in a multimodal model has been shown as an effective way to integrate available sources of information [87]. Here, we investigate combinations of this different information by integrating features from the inference models mentioned above in λ_i of (1). For example, λ_i in a combined model consisting of relatedness from text analysis and trust from probabilistic models would be a product of the provided comment's relatedness multiple by the user's gained reliability score.

IV. Evaluation

Here, we first overview the tool used in the current study. Then, general information about the data collected and the experimental settings for evaluating the consensus inference models are provided.[2] Finally, Section IV-C outlines a brief description of the evaluation metrics used for the analysis.

A. Tool

RiPPLE is an educational system that employs learnersourcing to create the resource repository. Fig. 3 uses screenshots of the platform to demonstrate some of its main functionality.

Fig. 3(a) illustrates an example of the interface used for creating learning resources. The example provided in the figure shows the page used for creating multiple-answer questions. RiPPLE relies on a peer-review process where students review and evaluate existing resources. RiPPLE assigns each resource to be evaluated by multiple moderators. Fig. 3(b) illustrates the interface used by a moderator for evaluating a resource using a rubric of four items, which asks moderators to rate a resource on alignment, correctness, difficulty level, and critical thinking encouragement [88]. Based on evaluations from the moderators, RiPPLE uses a consensus approach to infer resource quality. It currently uses an EM-inspired approach discussed in Section III-E. Fig. 3(c) shows an example of how evaluations and the inferred outcome are shared with the author, moderators, and instructors. The authors of the approved resources are encouraged to update their resources based on the feedback provided before they are added to the course repository. The authors of rejected resources can update and resubmit their resources. Approved resources are then used in RiPPLE, which is an adaptive educational system [9] at its core, to offer personalized learning by dynamically changing instructions tailored to the individual needs of students. Fig. 3(d) shows the personalized practice interface in RiPPLE. The upper part illustrates the learner model in the form of an interactive visualization widget that allows students to view an abstract representation of their knowledge state on a set of topics associated with a course offering. The lower part of the practice interface displays learning resources recommended to a student based on their learning needs using the recommender system outlined in [89].

B. Data Collection

The data used in this study are obtained from trialing RiPPLE in semesters 1 and 2 in 2020. Around 40 courses utilized RiPPLE in their offerings at The University of Queensland that year. However, in this work, we only gathered and reported data from the five offerings: Introduction to Information Systems in two semesters (code: INFS1-2), The Brain and Behavioural Sciences in two semesters (code: NEUR1-2), and Artificial Intelligence in semester 2 (code: COMP), that had the highest level of students participation and a considerable number of spot-checked resources from instructors for evaluation. General information about the datasets is presented in Table II. In total, 77 297 moderations were submitted on 12 803 resources by 2141 undergraduate students in these five courses' offerings. We applied an inclusion criterion for a meaningful and fair comparison [90] among different models where only resources with at least four moderations were selected for evaluation (e.g., 911 resources out of 2095 for INFS1). In addition, each course offering had a different number of instructors, including course coordinators, lecturers, and teaching assistants. In total, 28 instructors (7 from INFS1, 3 INFS2, 6 NEUR1, 7 NEUR2, and 5 COMP) spot-checked 694 resources in the selected datasets. These resources also received 4918 moderations from learners. These spot-checked resources are put aside as the test set for evaluation.

At the beginning of each semester, instructors determine the number of topics in each course. Then, authoring students would tag their created resources with one or more predefined

[2] Approval from our Human Research Ethics Committee #2018000125 was received for conducting the current study.

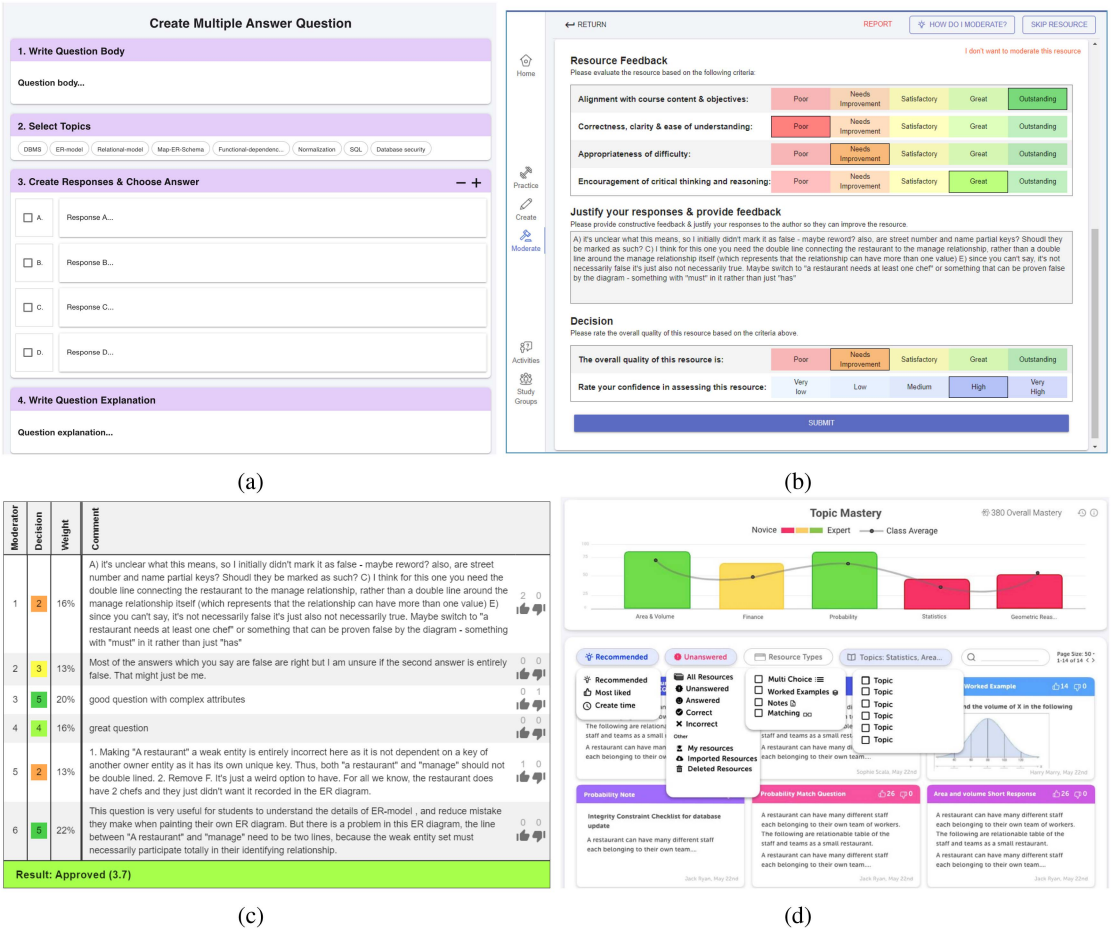

Fig. 3. Four of the main interfaces of RiPPLE for (a) resource creation, (b) evaluation rubric, (c) moderation feedback, and (d) personalized practice.

TABLE II
DATASET DETAILS FOR THE NUMBER OF RESOURCES, STUDENTS, AND MODERATIONS IN EACH COURSE

	INFS1			INFS2			NEUR1			NEUR2			COMP		
Data	Total	Selected	Spot-Checks	Total	Selected	Spot-Checks	Total	Selected	Spot-Checks	Total	Selected	Spot-Checks	Total	Selected	Spot-Checks
# Resources	2,095	911	112	1,835	921	41	4,875	4,851	145	2,803	2,757	303	1,195	926	93
# Students	389	378	250	385	372	127	532	526	304	535	527	483	300	295	191
# Moderations	6,991	4,327	508	6,182	4,167	165	28,152	28,090	728	30,642	30,559	3,131	5,330	4,912	386

topics. The resources were assigned to seven topics such as `Relational Models` for INFS, five topics such as `Brain Development` for NEUR, and five topics such as `Reasoning about other agents` for COMP. Students have submitted a total number of 47 338, 27 467, 62 540, 27 480, and 19 097 answers to the questions on the moderated resources for INFS1, INFS2, NEUR1, NEUR2, and COMP, respectively. Students' performance in each course and topic is used to measure their competence.

All courses employed a rubric in which students' participation in RiPPLE contributed 10% of their final grades. The grade associated with RiPPLE was conditional on students' participation in the moderation process but each course had a somewhat different requirement. For example, in NEUR, practicing on the platform and answering resources was not required while students were expected to moderate more than twice as many resources as INFS and COMP, which may have contributed to a differing moderation behavior. More specifically, there were 4 rounds of assessments in INFS offerings, and students were required to answer at least 10 resources correctly, create at least 1 resource, and moderate 4 or more resources on any topic in each round. In NEUR offerings, there were 5 rounds of assessments and students were required to create at least 1 resource and submit at least 10 moderations in each round. In COMP, students were required to answer at least 8 resources, create 1, and moderate 4 or more in 4 rounds of assessments. Fig. 4 offers further insights from the five courses. Fig. 4(a) demonstrates that students in all courses generally provided a positive high rating to their peer's work. Fig. 4(b) and (c) demonstrates that learners had quite diverse behaviors regarding their moderation numbers and average comment lengths. In particular, it can be seen that students

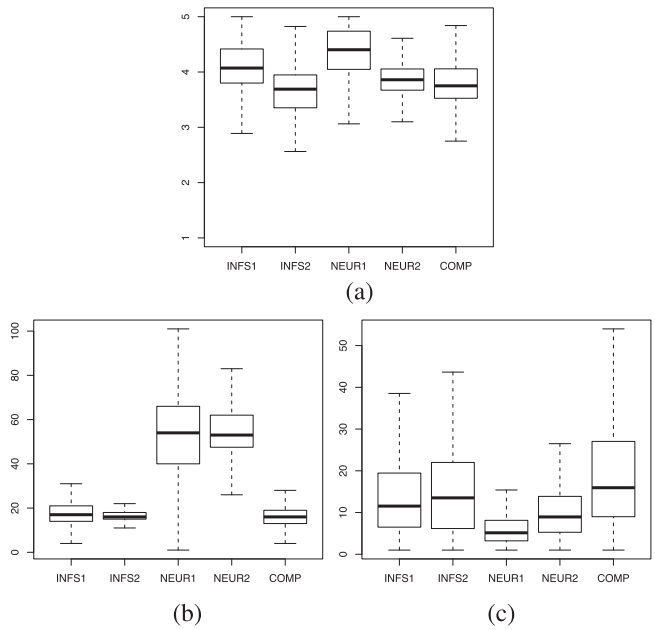

Fig. 4. Visualizations of the dataset comparing learners' behaviors in terms of (a) average rating, (b) the number of moderations, and (c) average length of comment in words.

from the NEUR course offerings conducted significantly more evaluations and wrote significantly shorter comments compared to INFS and COMP courses.

C. Metrics

The moderation process aims to decide whether or not a new resource created by a learner is good enough to be added to the repository of approved resources. Therefore, we perform our evaluation at a binary level by taking into account the threshold used by the system ($\Gamma = 3$) as the minimum required inferred rating for a resource to be approved. Accordingly, for both inferred ratings and instructor ratings, values of greater than or equal to three are put in binary class 1 and are categorized as "approved" and values of less than 3 are put in binary class 0 and are categorized as "Rejected." We use true positive rate (TPR), true negative rate (TNR), Area Under the Curve (AUC), and Accuracy (ACC) to report the performance of each of the models.

1) *TPR*, also known as sensitivity, which is computed based on the number of true positives (TP) (instructor approved and the inferred decision was also to approve) and false negatives (FN) (instructor approved while the inferred decision was to reject) as $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$. It shows the proportion of approved cases that were correctly classified.
2) *TNR*, also known as specificity, which is computed based on the number of true negatives (TN) (instructor rejected and the inferred decision was to reject) and false positive (FP) (instructor rejected while the inferred decision was to approve) as $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$. It shows the proportion of rejected cases that were correctly classified.
3) *AUC*, which is considered a reliable metric of aggregated classification performance that considers both TPR and TNR and is recognized as a suitable metric for evaluating labels with skewed distributions [91].
4) *ACC*, accuracy, which is computed based on the number of correct cases (i.e., TN and TP) and the total number of cases as $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$. It shows the proportion of total cases, which were correctly classified.

The absence of ground truth makes the selection of hyperparameters generally challenging for unsupervised learning problems [92]. Here, the required hyperparameters for probabilistic and text analysis models were learned via grid search, and cross-validation [93] in INFS1. Then, the same learned hyperparameters were used for other datasets. During this experiment, we observed that fine-tuning hyperparameters within individual models (e.g., selecting the standard deviation σ of Gaussian function in EM) makes no significant changes compared to the different feature selections (e.g., length over sentiment in the text analysis model).

V. Results

Table III shows the performance of the proposed consensus inference models in terms of TPR, TNR, AUC, and ACC. Results are categorized based on consensus models into summary statistics (cf. Section III-C), learner competence (cf. Section III-D), probabilistic (cf. Section III-E), text analysis (cf. Section III-F), and combined models (cf. Section III-G).

RQ1. Summary statistics. These approaches only use the student-provided numerical ratings for moderation. Summary statistics are commonly used in various applications for consensus; however, results reported in Table III show they perform poorly in the evaluation of SGC. These baselines (majority vote, mean, median, and debiased mean) generally have the lowest AUC among all the models.

RQ2. Learner competence. Incorporating course- and topic-based user competence models achieved improved AUC values compared to the baselines. The increase in the TNR values compared with baselines confirms that learner competency features can better identify moderators who have accurately labeled the poor-quality resources. Interestingly, the use of students' Elo rating and self-assessment of confidence has led to contradictory outcomes, leading to improvement in some courses like INFS1 while worsening the results in the other offerings.

RQ3. Probabilistic. Results show that reweighting students' contribution in the EM model based on how well their rating aligns with the inferred rating moderately improves the TNR values compared to the baselines. This model is the currently implemented consensus approach in RiPPLE, which only considers the numerical decision ratings from moderators. In contrast, the graph model achieved impressive improvements in TNRs compared to baselines with a slight decrease in TPRs. The MF approach predicts experts' behavior in SGC quality estimations based on the history of interactions between users and resources. Results show a considerable improvement in

TABLE III
COMPARISON BETWEEN THE INFERRED CONSENSUS RATINGS AND THE INSTRUCTORS' DECISIONS EVALUATED WITH TPR, TNR, AUC, AND ACC FOR MODERATION DECISIONS

Model	INFS1				INFS2				NEUR1				NEUR2				COMP			
	TPR	TNR	AUC	ACC	TPR	TNR	AUC	ACC	TPR	TNR	AUC	ACC	TPR	TNR	AUC	ACC	TPR	TNR	AUC	ACC
Summary Statistics:																				
Majority Vote	.96	.15	.56	.68	.94	.04	.49	.44	.99	.06	.53	.68	*1.00*	.02	.51	.63	.98	.02	.50	.52
Mean	.97	.21	.59	.71	.94	.09	.52	.46	*1.00*	.08	.54	.70	*1.00*	.03	.52	.63	*1.00*	.02	.51	.53
Median	*.99*	.15	.57	.70	.94	.04	.49	.44	*1.00*	.08	.54	.70	*1.00*	.00	.50	.62	*1.00*	.00	.50	.52
Debiased Mean	.95	.15	.55	.67	.94	.13	.54	.49	*1.00*	.08	.54	.70	*1.00*	.03	.51	.63	.98	.07	.52	.54
Learner Competence:																				
Course Correctness	.97	.18	.58	.70	.94	.17	.56	.51	*1.00*	.21	.60	.74	.99	.10	.55	.65	.94	.13	.54	.55
Topic Correctness	.97	.18	.58	.70	.94	.22	.58	.54	*1.00*	.17	.58	.72	.99	.10	.55	.66	.92	.13	.52	.54
User Elo Rating	.95	.31	.63	.72	.89	.09	.49	.44	*1.00*	.10	.55	.70	*1.00*	.04	.52	.64	.96	.13	.55	.56
User Confidence	.96	.28	.62	.72	*1.00*	.09	.54	.49	*1.00*	.06	.53	.69	*1.00*	.03	.52	.63	.98	.04	.51	.53
Probabilistic:																				
Expectation-Maximisation–EM	.95	.33	.64	.73	.94	.17	.56	.51	*1.00*	.10	.55	.70	.99	.03	.51	.63	.96	.13	.55	.56
Graph Trust Propagation	.81	.62	.71	.74	.94	.57	.75	.73	.88	.52	.70	.76	.94	.43	.68	.75	.90	.22	.56	.57
Matrix Factorisation–MF	.78	.54	.66	.70	.72	*1.00*	**.86**	**.88**	.80	.50	.65	.70	.55	*.68*	.62	.60	.31	*.67*	.49	.48
Text Analysis:																				
Length	.90	.56	.73	.79	.89	.26	.57	.54	.97	.35	.66	.77	.99	.26	.63	.71	.94	.16	.55	.56
Sentiment-Alignment	.90	.41	.66	.73	.89	.17	.53	.49	*1.00*	.12	.56	.71	*1.00*	.07	.53	.65	.90	.20	.55	.56
Similarity–GLEU	.93	.49	.71	.78	.94	.22	.58	.54	.95	.25	.60	.72	.99	.13	.56	.66	.96	.18	.57	.58
Relatedness–BERT	.93	.56	.75	**.80**	.94	.13	.54	.49	.98	.19	.58	.72	.99	.22	.60	.70	.92	.20	.56	.57
Combined models:																				
Length × Similarity	.90	.59	.75	.79	.94	.30	.62	.59	.93	.33	.63	.73	.98	.27	.63	.71	.94	.20	.57	.58
Length × Relatedness	.90	.62	**.76**	**.80**	.83	.39	.61	.59	.97	.40	.68	**.78**	.98	.31	.65	.73	.92	.20	.56	.57
Graph × Length × Relatedness	.86	*.67*	**.76**	.79	.83	.43	.63	.61	.86	*.56*	**.71**	.76	.91	.50	**.71**	**.76**	.90	.27	**.58**	**.59**

Numbers in bold highlight the best-gained results for AUC and ACC.

terms of identifying the TN cases compared to the previous models, whereas MF achieved the best outcome in INFS2. However, the TPR significantly decreased in the datasets with many discrepancies between students' and instructors' decisions. For example, in 47 cases in NEUR2 and 30 in COMP, all student-moderators approved a resource but instructors rejected it. Subsequently, the TNR dramatically increased from 0.03 in EM to 0.68 in MF for NEUR2 and from 0.13 to 0.67 for COMP with the cost of much sacrifice in TPR (from 0.99 to 0.55).

RQ4. Text analysis. Incorporation of linguistic information from comments provided as feedback to created resources into the text analysis models consistently improved the TNR, which has led to improvements in AUC values. Results also show that sentiment-alignment alone cannot indicate critical thinking well and subsequently has not further improved moderation outcomes than the length alone model. Comments should also mention aspects related to the resource under review to motivate more insightful feedback. The improvements in the last two text analysis models suggest an association between the similarity and relatedness of provided comments to the underreview resources and the quality of evaluations. Our findings suggest that extra features from text analysis of the provided comments help improve inferring the reliability of reviews.

RQ5. Combined models. Features from the models mentioned above can be combined in a wide variety of ways. For example, examining simply 3 models out of the 10 (3 Competence, 3 Probabilistic, and 4 Text) resulted in 120 models. While we exhaustively examined all conceivable combinations of two and three, only those top-performing models were listed that surpassed the results of single models. Equal weights for each model were employed in the combined models due to the unsupervised nature of the method and the observation that hyperparameter fine-tuning did not significantly influence the individual models' outcomes. The first two models, combining the length with similarity and relatedness, have significantly increased TNR and improved the AUC compared to the length alone model. A possible explanation for this might be that these models address a major limitation of using the length of a comment as a proxy of its effectiveness, which may lead to awarding long but uninformative comments. The third model uses a combination of features from the three most well-performing models—trust propagation, length, and relatedness, which led to significant growth in the TNR and identifying the poor-quality resources at the expense of moderately reducing TPR (identifying the high-quality resources) values. It improved TNRs by an average of 38% compared to the closest model in baselines across five courses. This substantial improvement also enhanced AUCs by an average of 14%.

Table III shows that the ACC has been improved in the combined models too. However, results also confirm that ACC is not a sensitive metric for skewed data. As a case in point, the Mean model presents an ACC of 0.70 in NEUR1 while it only has a TNR of 0.08. These results demonstrate that while most baselines overlook poor-quality resources, their relatively high ACC values might be misleading, resulting from a higher portion of the positive class (i.e., approved resources).

Fig. 5 drills down into the outcomes of the highest achieving model of each five categories of consensus approach on NEUR2, which has the maximum number of instructors' contributions in the spot-checking process compared to the other courses. Fig. 5(a) illustrates 3131 student moderations on 303 resources that an instructor also moderated. It shows that 1939 student moderations were received on the 188 resources approved by instructors (i.e., provided a rating of 3, 4, or 5), where $1899 \simeq 98\%$ students also approved these positive cases (i.e., $\mathrm{TP} \gg \mathrm{FN}$). In contrast, from the 1192 student moderations received on the 115 resources rejected by instructors (i.e., provided a rating of 1 or 2), only $120 \simeq 10\%$ students also rejected these negative cases (i.e., $\mathrm{TN} \ll \mathrm{FP}$). Fig. 5(b)–(f) shows the performance of different inference models at the resource level. Fig. 5(b) demonstrates that a baseline that uses a basic aggregation approach such as mean

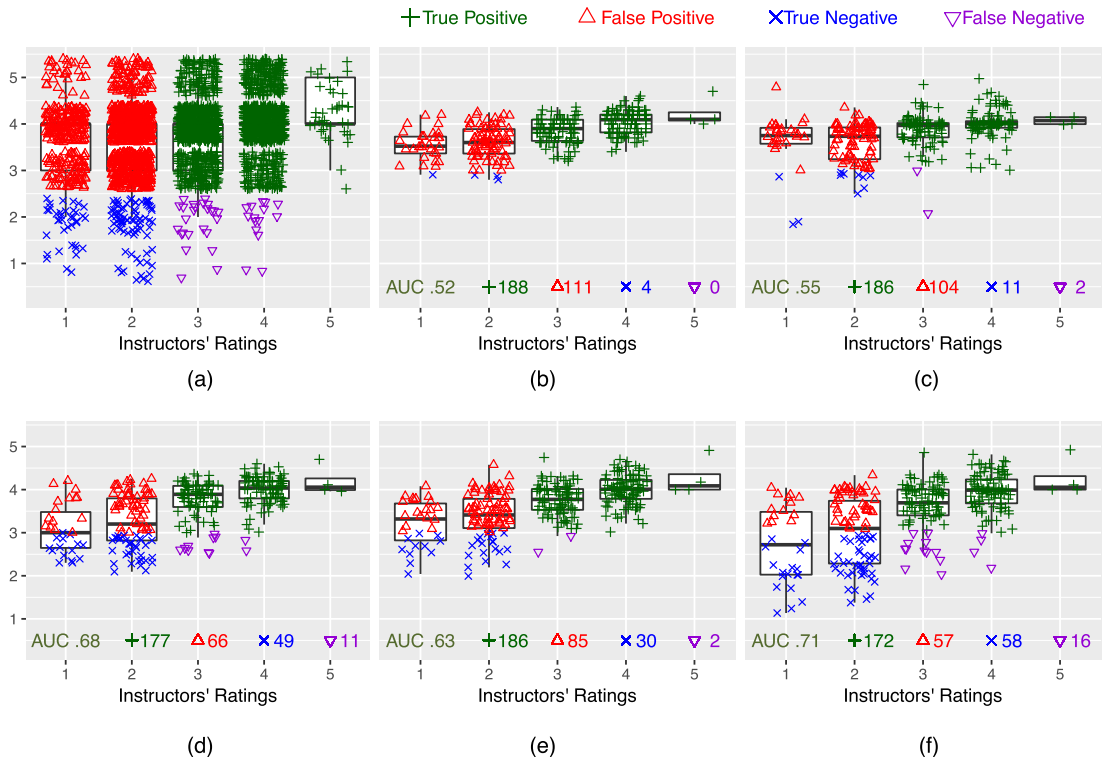

Fig. 5. Comparing the outcomes of inferring resources quality from (a) students' decision ratings, using different models: (b) Summary statistics (mean), (c) learner competence (course correctness), (d) probabilistic (graph trust propagation), (e) text analysis (length), and (f) a combined model (graph × length×relatedness) in NEUR2.

favors the majority. This model only rejected four resources and approved the rest, which indicates that the majority's decisions overthrew that of the minority wise-moderators. The competence model shown in Fig. 5(c) slightly enhanced the results by increasing the decision weights of the more qualified students in terms of correctly answering the resources. However, this improvement is not significant, and this model still failed to discern most of the poor-quality resources. Fig. 5 (d) and (e) shows that both probabilistic (graph) and text analysis (length) models led to better agreement with instructors' ratings. The probabilistic model identified the reliability of the moderator using a trust propagation approach in a network of students and instructors. The text analysis model identified the reliability of moderation by considering the length of a comment as a measure of effort. The combination of these two models took into account both the reliability of the moderation and the moderator. As shown in Fig. 5(f), the combination of the graph-based trust propagation from the probabilistic models with length and relatedness (BERT) from the text analysis models substantially distinguished poor-quality resources better than the commonly used baselines such as Mean (TN has increased from 4 to 58 cases). Despite the gained improvement, 57 FPs show that a considerable amount of poor-quality resources have not been identified. However, in 47 cases out of these 57 FPs, no student-moderator had rejected these resources. Therefore, most consensus approaches would not be able to identify these resources as poor quality. It is noteworthy that most of the inspected resources used as the evaluation set in this study are those complex cases in that the system could not reach a consensus with high confidence from students' decisions. In practice, RiPPLE flags and prioritizes a few resources to be inspected by instructors based on several criteria such as questionable distractors, low effectiveness based on student answers, reported resources, and high level of disagreements between student-moderators.

VI. Discussion

The results in Table III show that summary statistics are fragile against the skewed data, as evident by the high TPR values against the low TNRs and suffer from student evaluations that tend to overrate, resulting in approval of low-quality content. In learner competence models, results suggest a positive association between students' performance and their subjective sense of quality, which is in line with previous findings from the literature [94], [95]. We speculate Elo did not work well as the following:

1) students' ratings were not indicative of their ability;
2) this is because students might not have answered a sufficient number of questions;
3) the reason for not answering enough questions could be due to assessment criteria.

Incorporating self-assessment of confidence yielded inconsistent results, which are also in line with previous work that revealed contradictory findings regarding the reliability of self-assessment tests (e.g., [96], [97]).

Although EM from the probabilistic models has outperformed baselines in most cases, El Maarry et al. [54] argue that it might not be suitable for implementation in a domain with skewed data. They emphasize that using EM for consensus on data with long-tail distributions such as learner-sourced ratings will promote misbehaving in the system by strategic spammers who provide the most prevalent answer (e.g., a high rating here, cf. Section IV-B) in their evaluations. Nevertheless, our results also confirm that this model cannot suitably deal with the skewness of data and is still biased toward the majority who overrate. On the other hand, the finding suggests that the graph model can more effectively identify reliable and trustworthy student-moderators than the EM model. Also, the MF model was biased toward the skewed data and predicted a reject decision from instructors when students approved in most cases. The reason is that MF estimates each user's rating on an item considering their previous rating history in conjunction with all other users. Accordingly, when MF receives many cases with a set of students (i.e., $u_1, \ldots, u_{N-1}$) approving a resource, but the instructor (i.e., u_N) has rejected it, the probability of predicting a rejection from the instructor increase in the following evaluations.

The collected data show that the quality of the provided comments varies and can range from unhelpful comments such as "very good" and long positive feedback without much useful information to helpful comments such as "choice A is also correct." The proper use of comments for peer review can motivate learners to think more critically about the resources, which will also enhance their learning [98]. However, naively using features such as comment length in a live learning system can have the drawback of gaming the system by writing long but unhelpful comments. Employing the similarity and relatedness of feedback enhances the quality estimation and helps reduce the contribution of those strategic spammers in the review tasks. These features have also been heavily used outside education to estimate the quality and helpfulness of product reviews [29], [99].

In combined models, combinations of the features from the above models are considered. We have explored all possible combinations of two or three features from each subcategory of the five consensus approaches. However, we only report three top models that offered significant improvements in results. Results show that the length $\times$ relatedness of comments can better serve as a proxy for the amount of effort and critical thinking learners put into their evaluation. Finally, the last combined model (Graph $\times$ Length $\times$ Relatedness) was successful as it simultaneously takes into account both reliable reviewers by the trust propagation model and reliable reviews by the lengthy related comments. Although the considerable amount of FPs and a few borderline FN in Fig. 5 demonstrate an evident necessity for instructors' support during the moderation process, which is aligned with the findings of previous studies [11], [100], [101], the use of advanced consensus approaches can significantly reduce the instructors' load in the quality assessment of SGC at scale.

A. Limitations and Future Work

Although the results of this study suggest that student-moderations with quality textual feedback agree more with instructor-moderation, most moderations in the collected data consisted of short and unhelpful comments. One of the main problems might be offering students the same peer-review systems that experts use to provide feedback. The findings reveal a tie between the quality of comments in terms of length and relatedness and the quality of students' peer-review judgments. As a result, we hypothesize that more thoroughgoing comments will provoke students to be more critical, less lenient, and more confident in their reviews. A practical implication would be to develop a set of tips and a self-monitoring checklist for students to consider while writing their peer review and a set of quality control functions that automatically assess the quality of the submitted feedback and ask students to improve their inputs. We see the need for further research to establish methods to assist students as experts-in-training in providing more elaborative and compelling comments and evaluating the impacts of each method on students' behavior and the quality of peer review.

The simplicity of baseline approaches comes with a caveat inherent to student ratings: Students overrate poor quality resources, resulting in high aggregate scores and biassing correct moderation of poor quality resources. Also, the two main drawbacks of using learner competence models are 1) the scores might not exist for new users (cold start problem), and 2) learner interactions within the learning system (e.g., responding to a question) require the inference model to re-estimate user competency score at the moderation time and introduce additional computational expenses. Furthermore, combined models integrated various reliability estimates from individual inference models with equal weights to derive a more accurate final decision. However, it would be an interesting future path to study the impact of utilizing a validation set to identify ideal weights for each feature from different methods. The findings of this study suggest a tradeoff between consensus approaches performance and their level of complexity and explainability. In general, significant improvements in inferring SGC quality using the more complex approaches—probabilistic, text analysis, and combined models, make the final decisions more difficult to explain to students, which in turn lessens the accountability of the employed consensus for transparency [102]. A promising future direction is to develop consensus approaches that have high predictive performance but are still easy to explain.

VII. Conclusion

The peer-review process has been commonly used to evaluate the quality of SGC at scale. However, separating high-quality from low-quality resources is challenging as most students tend to be easy graders. The presented research aimed to examine the effectiveness of a range of consensus models in peer-reviewing SGC. In particular, we compared the performance of different models using students' subjective ratings, competency levels, reliabilities, textual feedback, and various combinations of these

features. The collected data in this study were from five-course offerings in two semesters, captured using a learnersourcing educational platform called RiPPLE. Our results suggest that the summary statistics such as majority vote, mean, and median, commonly used as the baselines in the redundancy-based strategies, cannot competently identify poor-quality resources. Although incorporating learners' competency improved the results compared to the baselines and demonstrated an association between student answering performance and evaluative judgment, it still fails to identify a large portion of negative cases. Interestingly, the estimation of reviewers' reliability using probabilistic models offered promising outcomes and enhanced SGC quality inference, especially in the graph-based trust propagation model. Moreover, text features such as length and relatedness proved to be good indicators of student-moderators effort and their reviews' reliability. Finally, our results showed that combining the probabilistic models as an indicator of reviewer reliability with the text analysis models as an indicator of review reliability achieved substantial improvement in discriminating between low- and high-quality SGC compared to the baselines, which can notably decrease the oversight workload of instructors.

REFERENCES

[1] J. Kim, "Learnersourcing: Improving learning with collective learner activity," Ph.D. dissertation, Dept. Elec. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2015.

[2] H. Khosravi, G. Demartini, S. Sadiq, and D. Gasevic, "Charting the design and analytics agenda of learnersourcing systems," in *Proc. 11th Int. Learn. Anal. Knowl. Conf.*, 2021, pp. 32–42.

[3] C. M. Reigeluth, B. J. Beatty, and R. D. Myers, *Instructional-Design Theories and Models, Volume IV: The Learner-Centered Paradigm of Education.* Evanston, IL, USA: Routledge, 2016.

[4] P. Denny, J. Hamer, A. Luxton-Reilly, and H. Purchase, "Peerwise: Students sharing their multiple choice questions," in *Proc. 4th Int. Workshop Comput. Educ. Res.*, 2008, pp. 51–58.

[5] H. Khosravi, S. Sadiq, and D. Gasevic, "Development and adoption of an adaptive learning system: Reflections and lessons learned," in *Proc. 51st ACM Tech. Symp. Comput. Sci. Educ.*, 2020, pp. 58–64.

[6] H. S. Alenezi and M. H. Faisal, "Utilizing crowdsourcing and machine learning in education: Literature review," *Educ. Inf. Technol.*, vol. 25, no. 4, pp. 2971–2986, 2020.

[7] S. Tackett et al., "Crowdsourcing for assessment items to support adaptive learning," *Med. Teacher*, vol. 40, no. 8, pp. 838–841, 2018.

[8] G. Gyamfi, B. Hanna, and H. Khosravi, "Supporting peer evaluation of student-generated content: A study of three approaches," *Assessment Eval. Higher Educ.*, vol. 47, no. 7, pp. 1129–1147, 2022.

[9] O.-C. Park and J. Lee, "Adaptive instructional systems," *Educ. Technol. Res. Develop.*, vol. 25, no. 1, pp. 651–684, 2003.

[10] J. L. Walsh, B. H. Harris, P. Denny, and P. Smith, "Formative student-authored question bank: Perceptions, question quality and association with summative performance," *Postgraduate Med. J.*, vol. 94, no. 1108, pp. 97–103, 2018.

[11] K. W. Galloway and S. Burns, "Doing it for themselves: Students creating a high quality peer-learning environment," *Chem. Educ. Res. Pract.*, vol. 16, no. 1, pp. 82–92, 2015.

[12] P. Denny, A. Luxton-Reilly, and B. Simon, "Quality of student contributed questions using peerwise," in *Proc. 11th Australas. Conf. Comput. Educ.—Vol. 95*, 2009, pp. 55–63.

[13] S. P. Bates, R. K. Galloway, J. Riise, and D. Homer, "Assessing the quality of a student-generated question repository," *Phys. Rev. Special Topics— Phys. Educ. Res.*, vol. 10, no. 2, 2014, Art. no. 020105.

[14] K. Cho and C. MacArthur, "Learning by reviewing," *J. Educ. Psychol.*, vol. 103, no. 1, 2011, Art. no. 73.

[15] Y. H. Cho and K. Cho, "Peer reviewers learn from giving comments," *Instructional Sci.*, vol. 39, no. 5, pp. 629–643, 2011.

[16] J. Tai, R. Ajjawi, D. Boud, P. Dawson, and E. Panadero, "Developing evaluative judgement: Enabling students to make decisions about the quality of work," *Higher Educ.*, vol. 76, no. 3, pp. 467–481, 2018.

[17] J. Whitehill, C. Aguerrebere, and B. Hylak, "Do learners know what's good for them? Crowdsourcing subjective ratings of OERs to predict learning gains," in *Proc. 12th Int. Conf. Edu. Data Mining*, 2019, pp. 462–467.

[18] A. Darvishi, H. Khosravi, and S. Sadiq, "Utilising learnersourcing to inform design loop adaptivity," in *Proc. Eur. Conf. Technol. Enhanced Learn.*, 2020, pp. 332–346.

[19] M. Petre et al., "Mapping the landscape of peer review in computing education research," in *Proc. Work. Group Rep. Innov. Technol. Comput. Sci. Educ.*, 2020, pp. 173–209.

[20] H. W. Marsh, U. W. Jayasinghe, and N. W. Bond, "Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability.," *Amer. Psychol.*, vol. 63, no. 3, 2008, Art. no. 160.

[21] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: Is the problem solved?," *Proc. VLDB Endowment*, vol. 10, no. 5, pp. 541–552, 2017.

[22] J. R. Wright, C. Thornton, and K. Leyton-Brown, "Mechanical TA: Partially automated high-stakes peer grading," in *Proc. 46th ACM Tech. Symp. Comput. Sci. Educ.*, 2015, pp. 96–101.

[23] H. Purchase and J. Hamer, "Peer-review in practice: Eight years of Aropä," *Assessment Eval. Higher Educ.*, vol. 43, no. 7, pp. 1146–1165, 2018.

[24] D. E. Paré and S. Joordens, "Peering into large lectures: Examining peer and expert mark agreement using peerscholar, an online peer assessment tool," *J. Comput. Assist. Learn.*, vol. 24, no. 6, pp. 526–540, 2008.

[25] D. K. Wind, R. M. Jørgensen, and S. L. Hansen, "Peer feedback with peergrade," in *Proc. 13th Int. Conf. e-Learn.,*2018, Art. no. 184.

[26] V. C. Raykar et al., "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, no. 4, pp. 1297–1322, 2010.

[27] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.

[28] H. Morise, S. Oyama, and M. Kurihara, "Bayesian probabilistic tensor factorization for recommendation and rating aggregation with multicriteria evaluation data," *Expert Syst. Appl.*, vol. 131, pp. 1–8, 2019.

[29] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 423–430.

[30] W. Xiong and D. Litman, "Automatically predicting peer-review helpfulness," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2011, pp. 502–507.

[31] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage.*, 2006, pp. 51–57.

[32] A. Darvishi, H. Khosravi, S. Abdi, S. Sadiq, and D. Gašević, "Incorporating training, self-monitoring and AI-assistance to improve peer feedback quality," in *Proc. 9th ACM Conf. Learn. Scale*, 2022, pp. 35–47.

[33] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1/2, pp. 1–39, 2010.

[34] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, 2009, pp. 324–331.

[35] M. P. McCurdy, W. Viechtbauer, A. M. Sklenar, A. N. Frankenstein, and E. D. Leshikar, "Theories of the generation effect and the impact of generation constraint: A meta-analytic review," *Psychon. Bull. Rev.*, vol. 27, no. 6, pp. 1139–1165, 2020.

[36] O. Chen, S. Kalyuga, and J. Sweller, "Relations between the worked example and generation effects on immediate and delayed tests," *Learn. Instruct.*, vol. 45, pp. 20–30, 2016.

[37] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham, "Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology," *Psychol. Sci. Public Int.*, vol. 14, no. 1, pp. 4–58, 2013.

[38] S. Moore, H. A. Nguyen, and J. Stamper, "Evaluating crowdsourcing and topic modeling in generating knowledge components from explanations," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2020, pp. 398–410.

[39] H. Khosravi, K. Kitto, and W. Joseph, "Ripple: A crowdsourced adaptive platform for recommendation of learning activities," *J. Learn. Anal.*, vol. 6, no. 3, pp. 91–105, 2019.

[40] E. L. Glassman, A. Lin, C. J. Cai, and R. C. Miller, "Learnersourcing personalized hints," in *Proc. 19th ACM Conf. Comput.-Supported Cooperative Work Social Comput.*, 2016, pp. 1626–1636.

[41] S. Weir, J. Kim, K. Z. Gajos, and R. C. Miller, "Learnersourcing subgoal labels for how-to videos," in *Proc. 18th ACM Conf. Comput. Supported Cooperative Work Social Comput.*, 2015, pp. 405–416.
[42] S. Bhatnagar, A. Zouaq, M. C. Desmarais, and E. Charles, "Learnersourcing quality assessment of explanations for peer instruction," in *Proc. Eur. Conf. Technol. Enhanced Learn.*, 2020, pp. 144–157.
[43] X. Wang, S. T. Talluri, C. Rose, and K. Koedinger, "Upgrade: Sourcing student open-ended solutions to create scalable learning opportunities," in *Proc. 6th ACM Conf. Learn. Scale*, 2019, pp. 1–10.
[44] P. J. Guo, J. M. Markel, and X. Zhang, "Learnersourcing at scale to overcome expert blind spots for introductory programming: A three-year deployment study on the python tutor website," in *Proc. 7th ACM Conf. Learn. Scale*, 2020, pp. 301–304.
[45] N. T. Heffernan et al., "The future of adaptive learning: Does the crowd hold the key?," *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 615–644, 2016.
[46] R. Bailey and M. Garner, "Is the feedback in higher education assessment worth the paper it is written on? Teachers' reflections on their practices," *Teach. Higher Educ.*, vol. 15, no. 2, pp. 187–198, 2010.
[47] D. Carless, "From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes," *Active Learn. Higher Educ.*, vol. 23, no. 2, pp. 143–153, 2022.
[48] N.-F. Liu and D. Carless, "Peer feedback: The learning element of peer assessment," *Teach. Higher Educ.*, vol. 11, no. 3, pp. 279–290, 2006.
[49] E. Er, Y. Dimitriadis, and D. Gašević, "A collaborative learning approach to dialogic peer feedback: A theoretical framework," *Assessment Eval. Higher Educ.*, vol. 46, no. 4, pp. 586–600, 2020.
[50] C. Mercader, G. Ion, and A. Díaz-Vicario, "Factors influencing students' peer feedback uptake: Instructional design matters," *Assessment Eval. Higher Educ.*, vol. 45, no. 8, pp. 1169–1180, 2020.
[51] L. De Alfaro and M. Shavlovsky, "Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments," in *Proc. 45th ACM Tech. Symp. Comput. Sci. Educ.*, 2014, pp. 415–420.
[52] V. Shnayder and D. C. Parkes, "Practical peer prediction for peer assessment," in *Proc. 4th AAAI Conf. Hum. Comput. Crowdsourcing*, 2016, pp. 199–208.
[53] B. K. Hassani, "The consensus approach," in *Scenario Analysis in Risk Management*. Berlin, Germany: Springer, 2016, pp. 39–50.
[54] K. El Maarry, U. Güntzer, and W.-T. Balke, "A majority of wrongs doesn't make it right-on crowdsourcing quality for skewed domain tasks," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, 2015, pp. 293–308.
[55] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1953–1961.
[56] T. Tian and J. Zhu, "Max-margin majority voting for learning from crowds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1621–1629.
[57] J. J. Williams et al., "Axis: Generating explanations at scale with learnersourcing and machine learning," in *Proc. 3rd ACM Conf. Learn. Scale*, 2016, pp. 379–388.
[58] S. Abdi, H. Khosravi, S. Sadiq, and G. Demartini, "Evaluating the quality of learning resources: A learnersourcing approach," *IEEE Trans. Learn. Technol.*, vol. 14, no. 1, pp. 81–92, Feb. 2021.
[59] A. Darvishi, H. Khosravi, and S. Sadiq, "Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach," in *Proc. 8th ACM Conf. Learn. Scale*, 2021, pp. 139–150.
[60] F. Saab, I. H. Elhajj, A. Kayssi, and A. Chehab, "Modelling cognitive bias in crowdsourcing systems," *Cogn. Syst. Res.*, vol. 58, pp. 1–18, 2019.
[61] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang, "Domain-weighted majority voting for crowdsourcing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 163–174, 2018.
[62] A. Darvishi, H. Khosravi, S. Sadiq, and B. Weber, "Neurophysiological measurements in higher education: A systematic literature review," *Int. J. Artif. Intell. Educ.*, vol. 32, pp. 413–453, 2022.
[63] S. Bull, "There are open learner models about!," *IEEE Trans. Learn. Technol.*, vol. 13, no. 2, pp. 425–448, Apr.–Jun. 2020.
[64] S. Abdi, H. Khosravi, S. Sadiq, and A. Darvishi, "Open learner models for multi-activity educational systems," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2021, pp. 11–17.
[65] R. Urena, G. Kou, Y. Dong, F. Chiclana, and E. Herrera-Viedma, "A review on trust propagation and opinion dynamics in social networks and group decision making frameworks," *Inf. Sci.*, vol. 478, pp. 461–475, 2019.
[66] G. Wang, S. Xie, B. Liu, and S. Y. Philip, "Review graph based online store review spammer detection," in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 1242–1247.
[67] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proc. 13th Int. Conf. World Wide Web*, 2004, pp. 403–412.
[68] J. Tu, G. Yu, C. Domeniconi, J. Wang, G. Xiao, and M. Guo, "Multi-label crowd consensus via joint matrix factorization," *Knowl. Inf. Syst.*, vol. 62, no. 4, pp. 1341–1369, 2020.
[69] H. Morise, S. Oyama, and M. Kurihara, "Collaborative filtering and rating aggregation based on multicriteria rating," in *Proc. IEEE Int. Conf. Big Data*, 2017, pp. 4417–4422.
[70] H. J. Jung and M. Lease, "Improving quality of crowdsourced labels via probabilistic matrix factorization," in *Proc. Workshops 26th AAAI Conf. Artif. Intell.*, 2012, pp. 101–106.
[71] W. Xiong and D. Litman, "Understanding differences in perceived peer-review helpfulness using natural language processing," in *Proc. 6th Workshop Innov. Use NLP Building Educ. Appl.*, 2011, pp. 10–19.
[72] D. Duret, R. Christley, P. Denny, and A. Senior, "Collaborative learning with peerwise," *Res. Learn. Technol.*, vol. 26, pp. 1–13, 2018.
[73] M. Fan, C. Feng, L. Guo, M. Sun, and P. Li, "Product-aware helpfulness prediction of online reviews," in *Proc. World Wide Web Conf.*, 2019, pp. 2715–2721.
[74] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[75] S. Xu, S. E. Barbosa, and D. Hong, "BERT feature based model for predicting the helpfulness scores of online customers reviews," in *Proc. Future Inf. Commun. Conf.*, 2020, pp. 270–281.
[76] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," 2019, *arXiv:1908.10084*.
[77] S. Chang, P. Dai, J. Chen, and E. H. Chi, "Got many labels? Deriving topic labels from multiple sources for social media posts using crowdsourcing and ensemble learning," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 397–406.
[78] J. Zhang, M. Wu, and V. S. Sheng, "Ensemble learning from crowds," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 8, pp. 1506–1519, Aug. 2018.
[79] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Proc. Explainable AI: Interpreting, Explaining, Visualizing Deep Learn.*, 2019, pp. 5–22.
[80] S. Krishnan, J. Patel, M. J. Franklin, and K. Goldberg, "A methodology for learning, analyzing, and mitigating social influence bias in recommender systems," in *Proc. 8th Conf. Recommender Syst.*, 2014, pp. 137–144.
[81] T. Rinker, "Sentimentr: Calculate text polarity sentiment. Version 2.4.0," 2018.
[82] M. Naldi, "A review of sentiment computation methods with R packages," 2019, *arXiv:1901.08319*.
[83] A. Darvishi, "Translation invariant approach for measuring similarity of signals," *J. Comput. Eng.*, vol. 1, pp. 21–29, 2009.
[84] H. Hassanpour and A. Darvishi, "A geometric view of similarity measures in data mining," *Int. J. Eng.*, vol. 28, no. 12, pp. 1728–1737, 2015.
[85] H. Hassanpour, A. Darvishi, and A. Khalili, "A regression-based approach for measuring similarity in discrete signals," *Int. J. Electron.*, vol. 98, no. 9, pp. 1141–1156, 2011.
[86] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.
[87] T. Damoulas and M. A. Girolami, "Combining feature spaces for classification," *Pattern Recognit.*, vol. 42, no. 11, pp. 2671–2683, 2009.
[88] G. Gyamfi, B. E. Hanna, and H. Khosravi, "The effects of rubrics on evaluative judgement: A randomised controlled experiment," *Assessment Eval. Higher Educ.*, vol. 47, no. 1, pp. 126–143, 2022.
[89] H. Khosravi, K. Cooper, and K. Kitto, "Riple: Recommendation in peer-learning environments based on knowledge gaps and interests," *J. Educ. Data Mining*, vol. 9, no. 1, pp. 42–67, 2017.
[90] Y. Li, B. I. P. Rubinstein, and T. Cohn, "Truth inference at scale: A Bayesian model for adjudicating highly redundant crowd annotations," in *Proc. World Wide Web Conf.*, 2019, pp. 1028–1038.
[91] C. Ferri, J. Hernández-Orallo, and P. A. Flach, "A coherent interpretation of AUC as a measure of aggregated classification performance," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 657–664.
[92] X. Fan, Y. Yue, P. Sarkar, and Y. R. Wang, "On hyperparameter tuning in general clustering problems," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2996–3007.

[93] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 281–305, 2012.
[94] R. A. Bjork, J. Dunlosky, and N. Kornell, "Self-regulated learning: Beliefs, techniques, and illusions," *Annu. Rev. Psychol.*, vol. 64, pp. 417–444, 2013.
[95] L. P. Macfadyen, S. Dawson, S. Prest, and D. Gašević, "Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations," *Assessment Eval. Higher Educ.*, vol. 41, no. 6, pp. 821–839, 2016.
[96] J. A. Ross, "The reliability, validity, and utility of self-assessment," *Practical Assessment, Res., Eval.*, vol. 11, no. 1, 2006, Art. no. 10.
[97] E. Panadero, G. T. Brown, and J.-W. Strijbos, "The future of student self-assessment: A review of known unknowns and potential directions," *Educ. Psychol. Rev.*, vol. 28, no. 4, pp. 803–830, 2016.
[98] D. Carless and D. Boud, "The development of student feedback literacy: Enabling uptake of feedback," *Assessment Eval. Higher Educ.*, vol. 43, no. 8, pp. 1315–1325, 2018.
[99] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proc. Joint Conf. EMNLP-CoNLL*, 2007, pp. 334–342.
[100] W. Wang, B. An, and Y. Jiang, "Optimal spot-checking for improving evaluation accuracy of peer grading systems," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 833–840.
[101] A. Darvishi, H. Khosravi, S. Sadiq, and D. Gašević, "Incorporating AI and learning analytics to build trustworthy peer assessment systems," *Brit. J. Educ. Technol.*, vol. 53, pp. 844–875, 2022.
[102] H. Khosravi et al., "Explainable artificial intelligence in education," *Comput. Educ.: Artif. Intell.*, vol. 3, 2022, Art. no. 100074.

Ali Darvishi received the B.Sc. and M.Sc. degrees in electrical engineering and electronics from the University of Mazandaran, Babolsar, Iran, in 2006 and 2009, respectively. He is currently working toward the Ph.D. degree in computer science with the Data Science group from School of Information Technology and Electrical Engineering and Institute for Teaching and Learning Innovation, The University of Queensland, Brisbane, QLD, Australia.

His research interests focus on employing learning analytics, machine learning, and data science to enhance learning and learning experiences, especially in evaluating and improving peer feedback quality.

Hassan Khosravi received the Ph.D. degree in computer science from Simon Fraser University, Burnaby, BC, Canada, in 2012.

He is currently an Associate Professor with the Institute for Teaching and Learning Innovation and an Affiliate Academic with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD, Australia. In his research, he draws on theoretical insights from the learning sciences and exemplary techniques from the fields of human–computer interaction, learning analytics and crowdsourcing to design, implement, validate, and deliver sociotechnical solutions that contribute to the delivery of learner-centered, data-driven learning at scale.

Afshin Rahimi received the B.S. degree in computer science and the M.S. degree in computational linguistics from the Sharif University of Technology, Tehran, Iran, in 2006 and 2013, respectively, and the Ph.D. degree in computer science from The University of Melbourne, Melbourne, VIC, Australia, in 2018.

He is currently a Lecturer with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD, Australia. His research interests fall within the fields of natural language processing, social network analysis, and machine learning.

Shazia Sadiq received the master's degree in computer science from the Asian Institute of Technology, Bangkok, Thailand, and the Ph.D. degree in information systems from The University of Queensland, Brisbane, QLD, Australia, in 1993 and 2002, respectively.

She is currently a Professor with the School of Information Technology and Electrical Engineering, The University of Queensland. She is part of the Data Science research group and is involved in teaching and research in databases and information systems. Her main research interests include innovative solutions for business information systems that span several areas, including business process management, governance, risk and compliance, data quality management, workflow systems, and service science.

Dragan Gašević received the master's and Ph.D. degrees in computer science from the University of Belgrade, Belgrade, Serbia, in 2002 and 2005, respectively.

He is currently a Professor of Learning Analytics with the Faculty of Information Technology and the Director of the Centre for Learning Analytics, Monash University, Melbourne, VIC, Australia. He is a Founder and was the President of the Society for Learning Analytics Research from 2015 to 2017. He is also an Honorary Professor with the School of Informatics, The University of Edinburgh, Edinburgh, U.K. His research interests include learning analytics involving developing computational methods that can shape next-generation learning technologies and advance our understanding of self-regulated and collaborative learning.