Credibility Metrics for Student-Assigned Labels of Textual Comments

Varsha Rao Akinepalli
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
vrakinep@ncsu.edu

Sourabh Pardeshi
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
sjpardes@ncsu.edu

Dhruv Mukesh Patel
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
dpatel49@ncsu.edu

Arnab Datta
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
adatta4@ncsu.edu

Banpreet Singh Chhabra
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
bchhabr@ncsu.edu

Edward F. Gehringer
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
efg@ncsu.edu

***Abstract*—This research paper describes how pattern recognition can be used to validate student-assigned labels and improve a training dataset natural language processing. In our preceding research, we attempted to enhance peer assessment by harnessing the power of natural language processing (NLP) and machine learning (ML) to critically evaluate the substance and quality of peer-review comments. This required training data, which we had students create by labeling review comments they had received from peers on various aspects of their work. Our previous paper delineated strategies for the automatic validation of labels ("tags") applied by students, aiming to enhance the reliability of the data fueling our ML algorithms. This paper examines the metrics introduced in our previous work, and studies how effective they are in the evaluation of actual labels assigned by students.**

***Index Terms*—Quality Control, Expertiza, Labeling, Krippendorf's alpha, Pattern Detection, Quality Control Metrics , Taggers, Tags, Reliable**

I. Introduction

Machine learning requires labeled data to train a model. Obtaining the labeled data is often difficult or expensive. In many cases, it is a good idea to crowdsource [1, 11] the labeling of these datasets. Our goal is to measure the quality of reviews. We use machine learning to find suggestions [25], detect problem statements [23], and assess helpfulness [3]. For this we need data that is labeled accurately. We assign students to label the reviews of their own work, because this distributes the labeling task widely, and because the authors of the work are best qualified to judge the utility of the reviews.

In more detail, the workflow proceeds as follows. Students, usually working as a team, submit a project. Other students are assigned to review the project, based on a rubric containing a set of criteria. Filling out the rubric involves assigning scores and, optionally, making comments on each criterion. To automatically rate the quality of the review, we would like to know how many of the reviewer comments contained suggestions, identified problems, or were (subjectively) considered helpful by the author. To train a machine-learning model to recognize these characteristics, we can use labels assigned by the student authors whose work has been reviewed. They are an ideal group to assess the review comments, because the comments are about *their* work, and they need to assimilate them to improve their final product.

It did not take long for us to recognize that not all of the student authors were careful about assigning labels (see Figure 1). Many of them whizzed along, assigning each label in less than one second. Strings of consecutive yeses and nos were common. For a time, we tried to manually assess the quality of the labeling. But this was labor intensive, and we could not guarantee the consistency, because assessments were made by multiple staff members and we did not have enough manpower to assign multiple raters to evaluate the same labels. We sought an automated strategy for assessing labeling, to go along with the automated strategy we were developing for assessing reviews. We outlined this approach in a paper [4] in last year's conference. It used three kinds of metrics: tagging speed (to detect labelers who were working "too fast"), detecting labelers whose labels disagreed with the labels assigned by others, and detecting repeated patterns in labeling (such as YYYNYYYNYYYN . . .). This paper reports on our experience in applying this strategy to actual peer assessments. This case study shows how we were able to obtain a reliable training dataset for machine learning, leveraging contributions from crowdworkers of uneven diligence. It should be useful to others who need training data for machine learning that their students can help them to construct.

Our first innovation was to combine all three of these metrics into a credibility score, a single number that could be used to assess the quality of the assigned labels. However, since this metric had not been validated in any way, we did not use it to assign scores for labeling. Instead we used our initial scoring metric, which awarded up to 10 points to each student for the labels they had assigned, and then subtracted

points for too-fast tagging and for too many consecutive Ys or Ns (this approach is described in greater detail below). A third metric is the pattern ratio, which calculates the fraction of labels that are found in repeated sequences of length one, two, or three.

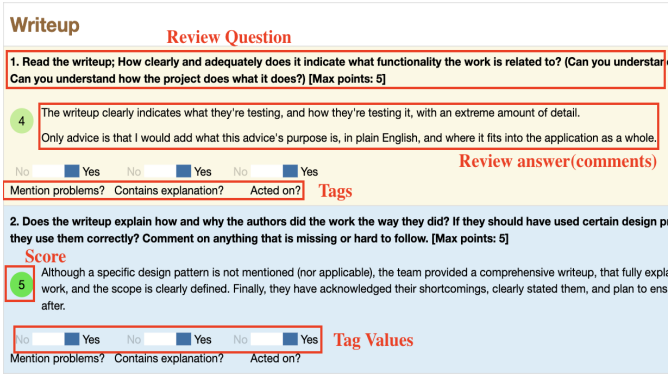

Fig. 1: Demonstration of tags and tag values

II. Related Work

The problem of vetting crowdsourced labels is not new. Several approaches have been developed for determining which labelers are to be believed, and for aggregating the work of many labelers into a a high-quality dataset.

Inter-rater reliability (IRR) is one common measure of the consensus among several evaluators who independently classify the same information. It has often been observed that discarding data with low consensus among raters enhanced the efficacy of machine-learning models trained with that data. A measure often used for this purpose is Krippendorff's alpha, which considers the actual agreement among raters and compares it with the amount of agreement expected by chance. One study applied Krippendorff's alpha to assess the consistency of sentence labels provided by Amazon Mechanical Turk workers [19], revealing its effectiveness in spotting subpar labels. Another investigation [16] found it useful for identifying poor-quality labels in user reviews labeled by students. A third study [20] introduced a weighted inter-rater reliability strategy to pinpoint low-performing annotators in video labeling tasks, which proved successful in reducing inconsistent labels.

Another method involves recognizing and discarding contributions from annotators who work unusually quickly. This technique, known as speed-based filtering, operates on the premise that rapid labelers might not thoroughly consider their tasks, potentially leading to substandard label quality. Research has been conducted to determine the impact of fast tagging on the quality of data labeled by a large group of people. Daniel et al. [6] developed a procedure detect rapid labelers using a median time threshold for labeling tasks, which, when applied, resulted in more accurate labels. Several other investigations showed that label quality and consistency improved when labels from too-fast annotators were removed [10, 12, 21], as studied among a group of Amazon Mechanical Turk workers. Other studies cautioned that fast tagging might not deliver high-quality labels, and suggested a combined approach with other quality-control methods for better results [13, 17].

Instead of trying to remove the contributions of low-quality labelers, we might instead construct an aggregation algorithm that accords greater weight to more reliable crowdworkers. This was the approach followed in Venzani et al's [22] entry in the 2013 CrowdScale—Shared Task Challenge for constructing reliable estimates of sentiments derived from a half-million tweets about the weather. The aggregation strategy has been followed in many subsequent research projects [2]. Many approaches have been proposed for combining labels from labelers of uneven quality into high-quality datasets [9, 18, 24]. That work, however, cannot be directly applied to our problem, because for any particular comment, we never have labels from more than 3 workers, and usually only one or two.

III. Methodology

A. Data Collection

Data was collected from our Expertiza [8] peer-assessment platform. We reviewed tags submitted by authors on peer reviews from several assignments. These assignments are created by the professor and teaching assistants, with students organized into teams for each task. Teams submit their work, which is then reviewed by individuals (who are not members of the submitting team). The reviews are based on rubrics, which ask reviewers to assign a score to the work based on each review criterion, and allow the reviewer to add a comment. Each team receives multiple reviews, and the students in these teams are responsible for tagging the comments they receive.

B. Toward metric definitions

This study aims to identify students who tag reliably and those who do not, based on peer assessment activities. To assess the reliability of taggers, we will evaluate them using three specific metrics. We have observed that reliable taggers tend to share certain characteristics: they do not tag too quickly, their tags generally align with their teammates' tags, and their tagging does not follow a discernible pattern. These observations yield the following metrics for tagging quality.

1) Time Taken to Tag: We use the timestamp information on each tag that is assigned to measure the intervals between tag assignments. Taggers who assign labels hastily may not adequately consider the review comment, leading to inconsistent or noisy labels. To identify too-fast taggers, we first calculate the time difference between consecutive tag timestamps, and apply a logarithmic transformation to these time differences to compress the distribution of timestamp values. These logarithmic timestamp differences are then summed to calculate the total time taken by each tagger to complete the assignment. The tagging speed is obtained by averaging the total time spent on all tags by each student. After analyzing these tagging speeds, a threshold is established. Students with an average tagging time above this threshold are likely taking sufficient time to thoughtfully assign tags. On the other hand,

those with an average tagging time below the threshold may be tagging hastily.

2) Krippendorff's alpha: When different students tag the same data, it's natural to see some variations in their responses. But if the taggers are working conscientiously, most of their tags they assign should agree. Inter-rater reliability (IRR) provides a measure to evaluate the degree of agreement among students who label the same data. Several measures exist for evaluating IRR, including Fleiss' kappa [7], Cohen's kappa [5], and Krippendorff's alpha [15].

Of the three mentioned above, Krippendorff's alpha is a prominent reliability coefficient designed to measure the level of agreement among multiple taggers. It is adaptable to various types of data, including nominal, ordinal, interval, ratio and and is capable of handling incomplete data. [14] . The Krippendorff's alpha values range from -1 to 1. An alpha value of -1 indicates no agreement, while a value of 1 represents perfect agreement.

α's general form is

$$\alpha = 1 - \frac{D_o}{D_e}$$

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck} \text{ metric} \delta^2_{ck}$$

$$D_e = \frac{1}{n(n-1)} \sum_c n_c \sum_k n_k \text{ metric} \delta^2_{ck}$$

The arguments in the two disagreements measures (D_o and D_e), o_{ck}, n_c, n_k and n, refer to the frequencies of values in coincidence matrices which is defined in the paper [14].

In this study, we have considered the Krippendorff's alpha values for all users within a team across various assignments. These alpha values indicate the degree of agreement among team members. An alpha value below 0 for a user suggests less agreement among that user and other team members, while values close to 1 indicate high agreement.

There are specific cases where Krippendorff's alpha cannot provide a reliable measure of inter-rater agreement, resulting in a value of None (NaN). The first condition occurs when a team consists of only one member. In such instances, there is no basis for calculating agreement or disagreement, causing Krippendorff's alpha to return NaN. The second condition arises when all students in a team assign the same tag to all data items, such as all "Yes" or all "No" tags, leading to no variability in the data. Lastly, Krippendorff's alpha also returns NaN when assignments have no tags assigned by any student.

3) Pattern Detection: Pattern detection is an effective method for identifying students who may be tagging in a repetitive manner, such as using a yes-no-yes sequence. Selecting answers randomly according to a predetermined pattern suggests that the student may not be paying adequate attention to the task. Such patterns indicate that the student is not engaging thoughtfully with the tagging process.

To implement this strategy, we developed an algorithm that looks for consecutive length-three and length- two patterns in all the tags of a student. This algorithm identifies and counts occurrences of a specific two and three-element pattern (e.g., YN, NNN, YNY) within a list of student tags. It then determines the maximum number of consecutive repetitions of this pattern. Initially, we examined patterns of longer lengths, but this approach did not help us distinguish between reliable and unreliable taggers. Upon focusing on students with multiple occurrences of longer patterns (e.g. 5- to 9-tag patterns), we observed that their tags were consistent and reliable, even though a pattern was repeated. We also recognized that it is generally challenging for an unreliable tagger to memorize and apply a pattern of five elements or more. Therefore, we limited our algorithm to considering shorter patterns.

C. Metric Information

We have devised several metrics to assess the reliability of labelers.

1) Points Deduction: The points deduction metric, as its name would indicate, gives the points deducted from a student's base score (which is based on the number of tags assigned). This metric takes into consideration both the time taken to assign tags and the occurrence of sequences comprising 10 or more consecutive "Yes" or "No" tags. The possible values for this metric range from 0 to 10. A score of 10 indicates that a student is tagging at a very rapid pace with numerous repetitions.

Let

- s be the tags set by the labeler
- t be the mean of the logarithms of the number of seconds between the setting of two consecutive tags,
- n be the number of tags in sequences of 10 or more consecutive "no"s, and
- y be the number of tags in sequences of 10 or more consecutive "yes"es.
- Let $w = (n + y)/s$.

Then

$$\text{PD} = (3 \text{ if } 0.95 < t \leq 1.3 \text{ or } 5 \text{ if } t < 0.95) + \left(3 \text{ if } 0.4 < w \leq \frac{2}{3} \text{ or } 5 \text{ if } w > \frac{2}{3}\right).$$

2) Credibility Score: The credibility score is designed to evaluate tagging behaviors by considering several key factors. Firstly, it takes into account the amount of time the tagger spends assigning a tag value, with a higher score indicating that the student took more time, which helps to identify hasty tagging behavior. Secondly, the score incorporates Krippendorff's alpha coefficient, which measures the level of agreement among students to identify contentious tags. Lastly, the score examines the frequency of "Yes" and "No" patterns in the tags.

Let

- ℓ be the time taken in seconds between the setting of two consecutive tags for each labeler (where "labeler" is a synonym for "tagger"),
- α be the Krippendorff's alpha coefficient for the labeler,

- r be the number of total repeating characters in all of the labeler's tags
- ℓ', r', and α' be the normalized values of ℓ, r, and α.
- Let $r'' = 1 - r'$, this would indicate that a labeler with the highest r value would have an r'' value of 0.

Then

$$\text{CS} = \frac{\ell' + \alpha' + r'}{3}$$

3) Pattern Ratio: The pattern ratio is a metric that evaluates each labeler by analyzing the presence of length-three, length-two, and alternating Yes and No patterns in their tags. Additionally, it considers the total number of repeated Yes and No responses. This metric aims to identify whether a student is memorizing specific tagging patterns and applying them consecutively in their tagging. Our analysis focuses on length-three patterns, as we believe this pattern length is the most likely to be remembered and utilized by students.

Let

- y be the total number of repeating Yes's in the labeler's tagging pattern,
- n be the total number of repeating No's in the labeler's tagging pattern,
- s be the number of tags that the labeler did set,
- m be the total number of consecutive length- two patterns in the labeler's tagging pattern (YN, NY),
- p be the maximum number of consecutive length-three patterns in the labeler's tagging pattern like YNY, NNY. Only values greater than 3 are considered.
- let $g = y + n + 2m + 3p$.

Then

$$\text{PR} = \frac{g}{s}.$$

IV. Results

A. Results for Experiment 1

The Data collected for this assignment consisted of 77 labelers, each tagging an average of 100 review comments. For this experiment,"helpful?" was the designated tag prompt (that is, the labelers were asked to decide whether or not each comment was helpful). Based on the feedback they received, labelers rated the usefulness of each review with a "Yes" or "No".

For each participant, the pattern ratio, credibility score, and points deduction were calculated. In order to enable cross-metric comparison, a regression analysis was conducted to examine the correlation between the pattern ratio and credibility score, as illustrated in Figure. 2.

The aforementioned results can be divided into *zones* to improve our understanding of the reliability of labelers, as shown in Figure 3. Zone 1 delineates the trustworthy zone, where labelers exhibit a high credibility score and a low pattern ratio. In contrast, Zone 2 delineates the untrustworthy labelers.

Figure 4 shows the frequency of points deducted in Experiment 1. As seen in the figure, only a small number of labelers had 3,5 or 10 points deducted. Since the majority of students did not have any points deducted, we use the "zones" as a more effective means of explaining student reliability.

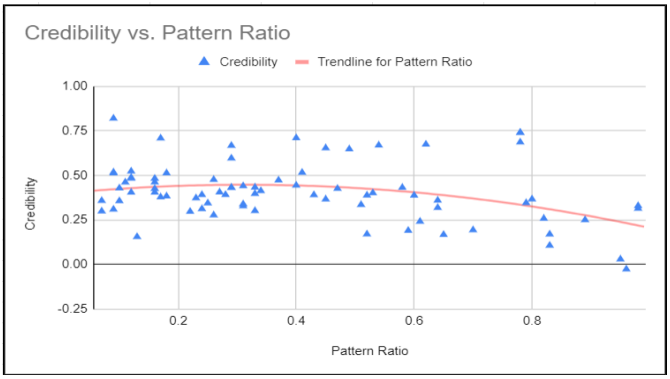

Fig. 2: Credibility Score Vs Pattern Ratio

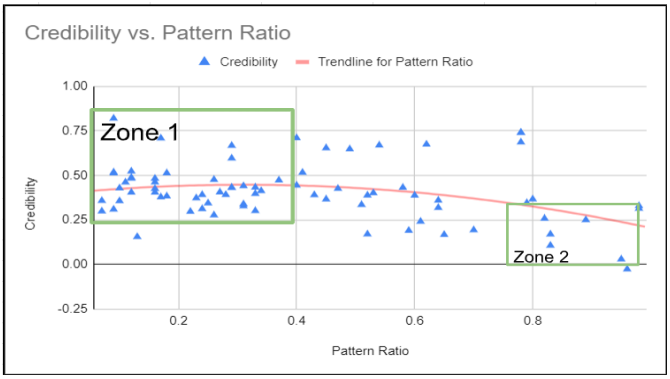

Fig. 3: Reliable and Unreliable Zones

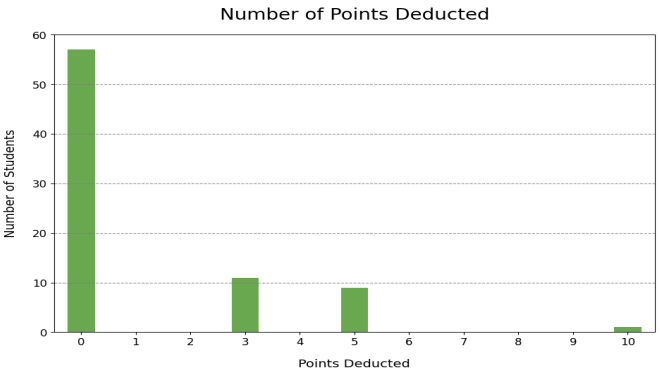

Fig. 4: Number of Points Deducted

We conducted a similar experiment on a different assignment, below are the results of that experiment.

B. Results for Experiment 2

This assignment comprised an average of 50 comments assessed by 48 labelers. The tag prompt used for this experiment again was "helpful?". Labelers assessed the helpfulness of reviews based on the comments received, tagging them as either "Yes" or "No".

The points deduction, credibility score, and pattern ratio were computed for all participants. To facilitate comparison across these metrics, a regression analysis was undertaken to examine the relationship between credibility score and pattern ratio, depicted in Figure 5.

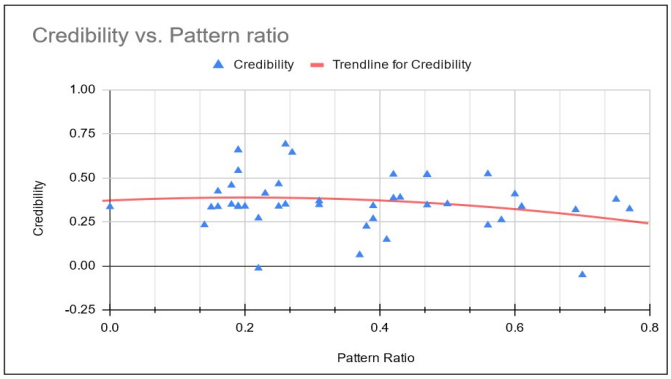

Fig. 5: Credibility Score Vs Pattern Ratio

To enhance our comprehension of the reliability of labelers, we can partition the above results into zones, as illustrated in Figure 6. Zone 1 delineates the reliable zone, characterized by labelers having a high credibility score and a low pattern ratio, while Zone 2 delineates the opposite scenario of unreliable labelers.

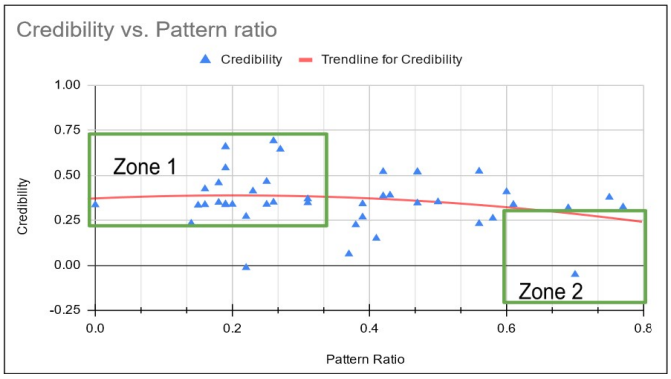

Fig. 6: Reliable and Unreliable Zones

Figure 7 illustrates the frequency of point deducted in Experiment 2. Consistent with our observations from Experiment 1, only a small number of labelers had any points deducted, which does not provide much guidance in determining reliability. Therefore, we once again use "zones" to provide a clearer explanation.

C. Summary

We conducted a ground-truth analysis to explore which metric yielded the most accurate results for identifying reliable taggers.

We analyzed the tagging behaviors of a randomly selected sample of students from the zones specified in Figures 3 and 6, as well as students outside these zones. The results indicated that some students demonstrated reliability in their

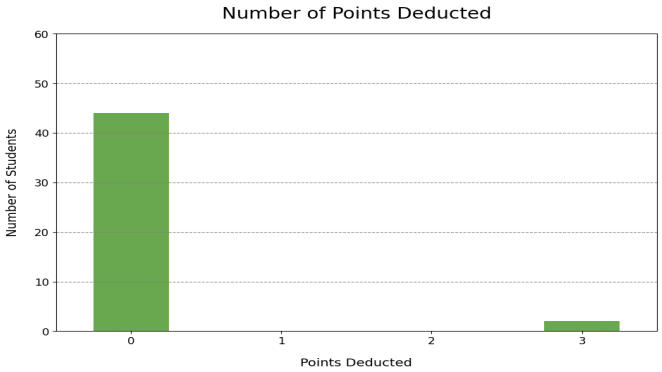

Fig. 7: Number of Points Deducted

tagging even if they were not part of the reliable zone. These students reviewed comments, the majority of which were helpful. However, their tendency to tag numerous comments with "Yes" inflated their pattern ratio, thereby pushing them out of the reliable zone.

We examined the ground truth tags of taggers in both the reliable and unreliable zones. Our observations revealed that students with higher credibility exhibited lower pattern ratios, and conversely, those with lower credibility had higher pattern ratios.

Our metrics provide a way to evaluate tagging objectively. By incorporating various factors such as tagging speed, team agreement, and patterns in tagging, this approach offers a well grounded assessment of a student tagger's reliability.

V. Future Scope

To ensure the robustness and generalizability of our metrics, additional experiments should be conducted across a broader range of assignments, academic fields, and curriculum levels. This would help to identify any potential biases or limitations in the current metrics and provide insights into how they can be adjusted or improved.

Incorporating large language models like GPT-4 for ground truth labeling could significantly enhance the accuracy of identifying reliable labelers. These models can provide a more nuanced understanding of the content and context of peer reviews, helping to distinguish between genuinely thoughtful feedback and superficial comments.

Acknowledgment

The authors express their gratitude to their colleagues for their contributions to this paper. Furthermore, we acknowledge the significant efforts of Dr. Edward Gehringer, whose research was reviewed and incorporated into our study.

References

[1] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 2334–2346, 2017.

[2] Sujoy Chatterjee, Anirban Mukhopadhyay, and Malay Bhattacharyya. A review of judgment analysis algorithms for crowdsourced opinions. *IEEE Transactions on Knowledge and Data Engineering*, 32(7):1234–1248, 2019.

[3] Ruixuan Shang Qinjin Jia Parvez Rashid Chengyuan Liu, Jialin Cui and Edward Gehringer. Generative ai for peer assessment helpfulness evaluation. In *Proceedings of the 17th International Conference on Educational Data Mining*, 2024.

[4] Banpreet Singh Chhabra and Edward F Gehringer. Quality control of crowd labeling for improving the quality of peer assessments. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2023.

[5] J Cohen. A coefficient of agreement for nomimal scales. *Educ Psychol Meas*, 20:2746, 1960.

[6] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40, 2018.

[7] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[8] Edward F Gehringer. Expertiza: Managing feedback in collaborative learning. In *Monitoring and assessment in online collaborative environments: Emergent computational technologies for e-learning support*, pages 75–96. IGI global, 2010.

[9] Mihai Georgescu and Xiaofei Zhu. Aggregation of crowdsourced labels based on worker history. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, pages 1–11, 2014.

[10] Eric Heim, Tobias Roß, Alexander Seitel, Keno März, Bram Stieltjes, Matthias Eisenmann, Johannes Lebert, Jasmin Metzger, Gregor Sommer, Alexander W Sauter, et al. Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*, 5(3):034002–034002, 2018.

[11] David R Karger, Sewoong Oh, and Devavrat Shah. Efficient crowdsourcing for multi-class labeling. In *Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, pages 81–92, 2013.

[12] Gabriella Kazai and Imed Zitouni. Quality management in crowdsourcing using gold judges behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 267–276, 2016.

[13] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318, 2013.

[14] K Krippendorff. Computing krippendorff's alpha-reliability. retrieved january 20, 2012, 2011.

[15] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70, 1970.

[16] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[17] Matthew Lease. On quality control and machine learning in crowdsourcing. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*, 2011.

[18] Yuan Li, Benjamin Rubinstein, and Trevor Cohn. Exploiting worker correlation for label aggregation in crowdsourcing. In *International conference on machine learning*, pages 3886–3895. PMLR, 2019.

[19] Rion Snow, Brendan O' connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.

[20] Anirudh Som, Sujeong Kim, Bladimir Lopez-Prado, Svati Dhamija, Nonye Alozie, and Amir Tamrakar. Automated student group collaboration assessment and recommendation system using individual role and behavioral cues. *Frontiers in Computer Science*, 3:728801, 2021.

[21] Paul Upchurch, Daniel Sedra, Andrew Mullen, Haym Hirsh, and Kavita Bala. Interactive consensus agreement games for labeling images. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, pages 239–248, 2016.

[22] Matteo Venanzi, John Guiver, Gabriella Kazai, and Pushmeet Kohli. Bayesian combination of crowd-based tweet sentiment analysis judgments. *Proc. Crowdscale Shared Task Challenge*, 3, 2013.

[23] Yunkai Xiao, Gabriel Zingle, Qinjin Jia, Harsh R Shah, Yi Zhang, Tianyi Li, Mohsin Karovaliya, Weixiang Zhao, Yang Song, Jie Ji, et al. Detecting problem statements in peer assessments. *arXiv preprint arXiv:2006.04532*, 2020.

[24] Jing Zhang, Victor S Sheng, and Jian Wu. Crowdsourced label aggregation using bilayer collaborative clustering. *IEEE transactions on neural networks and learning systems*, 30(10):3172–3185, 2019.

[25] Gabriel Zingle, Balaji Radhakrishnan, Yunkai Xiao, Edward Gehringer, Zhongcan Xiao, Ferry Pramudianto, Gauraang Khurana, and Ayush Arnav. Detecting suggestions in peer assessments. *International Educational Data Mining Society*, 2019.