International Journal of Education, Vocational and Social Science
https://e-journal.citakonsultindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

Discrepancies Between Star Ratings And Textual Sentiment In MOOC Reviews And Their Pedagogical Implications

Wiputra Cendana*
Universitas Pelita Harapan
Nur Indah Eriyathi
Universitas Pelita Harapan
Wiginy Kusliawan
Universitas Pelita Harapan
David Habsara Hareva
Universitas Pelita Harapan
Khoe Yao Tung
Universitas Pelita Harapan
Ranny Fransiska
Universitas Pelita Harapan

Corresponding Author: wiputra.cendana@uph.edu

Abstract

Star ratings summarize course evaluations but may not represent textual sentiment or criticism relevant to instructional improvement. This exploratory secondary study examines 107,018 MOOC reviews and two sets of human annotations for 120 reviews, with 40 reviews in each rating stratum: 1–2, 3, and 4–5 stars. Exact text matching, cross-tabulation, percentage agreement, and Cohen's kappa were used without changing annotators' labels. The corpus contained 100,038 distinct texts; 54 groups of identical texts had conflicting sentiment categories derived from ratings. Agreement between Annotators A and B was 94.17% for sentiment (κ = 0.9018), 87.50% for primary pedagogical aspect (κ = 0.8316), and 100% for the criticism indicator (κ = 1.0000). Agreement with rating proxies was only 70.83% for A and 69.17% for B. Among 40 three-star reviews, only 12 and 8 were labeled neutral by A and B, respectively. Sixty-six reviews contained criticism, including three of the 40 reviews with four or five stars. Textual evidence concerned assessment alignment, learning support, and access to materials. Label agreement

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

establishes neither indicator validity nor intervention effectiveness. Findings are descriptive for a sample with an undocumented selection mechanism; educational decisions require review of textual evidence and course context.

Keywords: MOOC reviews; star ratings; human annotation; coding agreement; pedagogical criticism; learning analytics.

INTRODUCTION

Reviews of massive open online courses (MOOCs) contain learners' evaluations of content, teaching, assignments, support, and their experience of using the platform. Star ratings facilitate summaries, but a single score can combine different experiences. Learners may appreciate the content while finding assignment rubrics unclear, or consider a course valuable while criticizing its payment policy. Overall evaluations, textual sentiment, and the presence of problems that could be addressed therefore provide related but distinct information.

Previous research has used text analysis to examine pedagogical factors in MOOCs. Li et al. (2022) analyzed reviews and distinguished key factors according to course orientation, while Awadh et al. (2025) reviewed approaches to aspect-based sentiment analysis in MOOCs. These developments indicate that applying sentiment analysis alone is no longer a novel contribution. A continuing issue is how readily available labels, particularly ratings, relate to interpretations of review content and criticism of learning experiences.

This study examines that relationship through a corpus audit and two existing sets of human annotations. It provides traceable descriptive evidence of discrepancies between ratings and sentiment, coding agreement across three dimensions, and criticism within reviews that are positive overall. The analysis does not treat either ratings or annotators as a single source of truth. Its purpose is to clarify what information is lost when learning experiences are reduced to global categories.

Research questions

RQ1. How closely do sentiment categories derived from ratings agree with the sentiment labels recorded in two annotation sets, particularly for three-star reviews?

RQ2. What level of coding agreement is observed for sentiment, primary pedagogical aspect, and the presence of actionable criticism?

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

RQ3. What pedagogical criticism can be traced to review texts, including reviews with positive ratings or sentiment, and what limits apply when using it to develop recommendations?

CONCEPTUAL BACKGROUND AND RELATED RESEARCH

Ratings and interpretations of review content

Ratings express an overall evaluation on the scale provided by a platform. Textual sentiment, in contrast, is an interpretation of written expressions. A discrepancy does not automatically indicate an error in either representation: a learner may evaluate the experience as a whole, whereas an annotator may prioritize the most prominent criticism in the text. A three-star rating may also reflect a mixed experience rather than an absence of polarity. Aspect-based analysis helps distinguish the objects of praise and criticism, allowing interpretation to extend beyond global sentiment (Awadh et al., 2025).

Coding agreement and pedagogical criticism

Cohen's κ adjusts the observed proportion of agreement for the agreement expected from the category distributions (Cohen, 1960). It should be interpreted alongside label distributions and case counts rather than as proof of construct validity (Sim & Wright, 2005). Coding guidelines, explanations of category boundaries, and examination of disagreements help establish whether differences arise from reviews addressing multiple aspects or from unclear decision rules (O'Connor & Joffe, 2020).

In education, criticism may be useful when it identifies features of learning that can be examined, such as alignment between videos and assignments, the clarity of instructions, or the availability of subtitles. Constructive alignment connects learning objectives, activities, and assessment (Biggs, 1996), while feedback research emphasizes information that helps determine steps for improvement (Hattie & Timperley, 2007). However, an "actionable" label still requires context and pedagogical judgment. The concept of an actionable insight itself requires an explanation of how information connects to possible action (Jørnø & Gynther, 2018).

Position of the present study

Li et al. (2022) examined factors contributing to MOOC quality, whereas the present study evaluates relationships between representations of evaluation at the review level. The methodological review by Awadh et al. (2025) situates this analysis within the need for

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

interpretation at the aspect level. This study connects label agreement, conflicts between categories derived from ratings, and evidence of criticism in the same examples so that pedagogical implications can be traced back to individual texts.

The relationship between evidence and action must also be distinguished from the measurement of impact. Wise and Jung (2019) examined how instructors use analytics through questions, interpretation, action, and assessment of impact. The present study examines evidence and develops implications. It does not measure the quality of instructors' decisions after using a dashboard.

METHODOLOGY

Design and unit of analysis

The study uses descriptive secondary analysis and examination of review content. The materials comprise a reviews.csv corpus containing 107,018 rows, two annotation CSV files from Annotators A and B, and a documentation workbook containing both annotation sets. Each corpus row has Id, Review, and Label fields, with ratings from 1 to 5. The unit of analysis is the recorded review rather than a unique learner, course, school, or learning outcome. Id is treated as the identifier of a source row.

The corpus analysis examines field completeness, rating distributions, distinct texts, and groups of identical texts with different ratings. Corpus statistics include all rows, without language filtering or removal of repeated texts. The data contain more than one language. The corpus audit contextualizes the coding results for the annotation sample.

Annotation sample and source matching

The annotation documentation contains 120 reviews with Code values R001–R120: 40 reviews with ratings of 1–2, 40 with a rating of 3, and 40 with ratings of 4–5. The mechanism and probability of selection within each stratum were unavailable for this secondary analysis. The sample is treated as a collection with equally sized strata and an unknown selection process, rather than as a representative random sample. No weighting was applied to estimate prevalence in the corpus.

The A and B records were aligned using Code, after which Review texts were checked for exact equality following UTF-8 CSV parsing. Texts were not lowercased, trimmed, or normalized for matching. Each text was located in the parent corpus, retaining all matching Id values and

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

ratings. Identical text appearing under multiple Id values was not treated as a single learner identity. The workbook was checked against each annotator's CSV file, and their contents were consistent.

Variables and label status

Rating proxy categories were defined as negative for ratings of 1–2, neutral for a rating of 3, and positive for ratings of 4–5. Three annotation variables were analyzed as recorded: Sentiment_Annotator, with three categories; Primary_Pedagogical_Aspect, with Content, Teaching, Assessment, Learning Support, Technical, and General categories; and Contains_Actionable_Criticism, with Yes or No categories. Notes were used to understand interpretive context rather than as a categorical variable in the calculation of κ.

The criticism indicator represents the annotators' recorded judgments about the presence of criticism they considered actionable. The analysis does not reassign these labels or claim that an action has proved effective. Annotator backgrounds, training, separation of the labeling processes, and access to ratings or other labels were not evaluated in this study. The results are therefore limited to the available coding agreement. The original A and B labels were retained; no consensus labels or new gold standard were created.

Descriptive analysis and verification of calculations

Observed agreement was calculated as the number of matching labels divided by the number of pairs. Cohen's κ was calculated separately for each nominal dimension and for agreement between each annotator's sentiment labels and the rating proxy (Cohen, 1960). Expected agreement was obtained by summing the products of the marginal proportions for each category. The formula is as follows.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

In this formula, P with subscript o is the observed proportion of agreement, and P with subscript e is the agreement expected from the category marginals. Label differences were tabulated by dimension and unique Code. Percentages are reported to two decimal places and κ to four decimal places; calculations used unrounded values.

Cross-tabulations compared the rating proxy with the sentiment labels of A and B. Criticism frequencies were calculated by rating stratum and within the positive sentiment labels assigned by each annotator. Because sample selection is undocumented, the statistics are

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

descriptive and do not support population estimates or tests of population differences. Pedagogical issues were interpreted using coded text examples traceable to source Id values. Recommendations are the authors' inferences from those examples and are not reported as intervention outcomes.

The supplement provides a Python script using the standard library, SHA-256 hashes of input files, a traceability table, label matrices, and aggregate results. The script does not modify reviews, ratings, or annotator labels. Annotation categories were not used to generate new predictions.

RESULTS

Corpus characteristics and quality

Table 1 Rating distribution in the corpus and annotation sample

Rating	Corpus n	Corpus %	Sample n
1	2,469	2.31	20
2	2,251	2.10	20
3	5,071	4.74	40
4	18,054	16.87	5
5	79,173	73.98	35
Total	107,018	100.00	120

No missing values were found in Id, Review, or Label. Ratings of 4–5 accounted for 97,227 of the 107,018 rows (90.85%), and the mean rating was 4.5811. In contrast, each proxy category accounted for one third of the annotation sample. This difference in composition means that sample frequencies cannot be used directly to describe the distribution of experiences in the corpus.

The corpus contained 100,038 distinct texts and 1,703 groups of texts appearing more than once. There were 6,980 additional occurrences beyond the first occurrence of each text. Ratings on the 1–5 scale differed within 402 groups; after mapping ratings into three proxy categories, 54 groups still contained conflicting categories, covering 1,152 rows. For example, a difference between ratings of 4 and 5 was not counted as a proxy conflict because both map to positive. Identical texts may come from different reviews, so these findings do not establish duplication of learners.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Non-English texts were observed, including at Id 551 and Id 6217, confirming that the corpus includes more than one language. These examples were not used to estimate language composition; no automated language identification was performed.

All 120 annotated texts were found in the corpus, and all had consistent source ratings. There were 122 matching rows: R007 matched Id 94539 and Id 104985, while R102 matched Id 17273 and Id 69855. Each pair shared the same text and rating. The traceability table retains all these Id values so that matching does not imply a single source identity where multiple matches exist.

Coding agreement

Table 2 Agreement between Annotators A and B on 120 reviews

Dimension	Agree n	Differ n	Agree %	κ
Sentiment	113	7	94.17	0.9018
Primary pedagogical aspect	105	15	87.50	0.8316
Presence of criticism	120	0	100.00	1.0000

Annotator A recorded 54 negative, 17 neutral, and 49 positive sentiment labels; Annotator B recorded 57 negative, 10 neutral, and 53 positive labels. All seven sentiment disagreements involved reviews labeled neutral by A and labeled negative in three cases and positive in four cases by B. There were no direct negative–positive disagreements. The complete matrix is presented in Appendix A.

Table 3 Distribution of primary pedagogical aspects in the sample

Aspect	Annotator A n	Annotator B n
Content	51	50
Teaching	20	23
Assessment	21	18
General	16	15
Learning Support	6	7
Technical	6	7

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

There were 15 disagreements on the primary aspect. One review, R119, differed on both sentiment and aspect, giving 21 unique Code values with at least one disagreement. Both annotators assigned Yes to 66 reviews and No to 54 reviews. Perfect agreement on the criticism indicator shows consistency in the decisions recorded for this sample; it does not demonstrate that the indicator's definition has been validated or that other annotators would reach the same decisions.

Agreement between ratings and textual sentiment

Table 4 Textual sentiment by rating stratum

Stratum	N	Neg. A	Neu. A	Pos. A	Neg. B	Neu. B	Pos. B
Rating 1–2	40	35	4	1	36	2	2
Rating 3	40	18	12	10	20	8	12
Rating 4–5	40	1	1	38	1	0	39

Neg., Neu., and Pos. denote negative, neutral, and positive, respectively. Annotator A agreed with the rating proxy on 85/120 reviews (70.83%; $\kappa = 0.5625$), while B agreed on 83/120 reviews (69.17%; $\kappa = 0.5375$). Thus, 35 reviews according to A and 37 according to B did not match the proxy. These figures describe discrepancies between representations rather than error rates for annotators or models.

The most pronounced differences occurred among three-star reviews: A labeled 18 negative, 12 neutral, and 10 positive; B labeled 20 negative, 8 neutral, and 12 positive. The neutral category derived from ratings did not match textual labels for 28/40 reviews according to A (70.00%) and 32/40 according to B (80.00%). This finding supports examining the content of three-star reviews in this sample, without implying that all three-star MOOC reviews are non-neutral.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Criticism in reviews with positive evaluations

Table 5 Presence of criticism by rating stratum

Stratum	n	Criticism Yes n	Criticism Yes %
Rating 1–2	40	32	80.00
Rating 3	40	31	77.50
Rating 4–5	40	3	7.50
Total sample	120	66	55.00

Criticism frequencies were identical for A and B in every stratum. Among reviews with ratings of 4–5, three reviews, coded R016, R036, and R116, still contained criticism. These cases demonstrate its presence in the sample; the proportion of 7.50% is not an estimate for all highly rated reviews in the corpus.

Positive textual evaluations can also coexist with criticism. Of the 49 reviews labeled positive by A, 10 contained criticism (20.41%); of the 53 labeled positive by B, 13 contained criticism (24.53%). The different denominators reflect each annotator's sentiment decisions. Filtering only negative reviews would therefore exclude some criticism already identified in the annotation documentation.

Traceability of pedagogical issues

Table 6 Examples of issues traced to source reviews

Code and Id	Rating	Aspect A / B	Issue paraphrase
R003 Id 97069	1	Assessment / Assessment	Lectures, assignments, and quizzes are misaligned; instructions and reading sources are unclear.
R005 Id 32112	2	Assessment / Technical	Links to quizzes are broken; the relevance of questions to the material is also criticized.
R036 Id 56277	5	Content / Content	The course is praised, but the learner requests advanced content and clearer explanations of notation on slides.
R066 Id 106237	1	Technical / Technical	English subtitles are reported to be available only for the first video.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

R084 Id 42172	3	Learning Support / Learning Support	Forum interaction is limited; the learner suggests encouragement to increase discussion.
R116 Id 40969	4	Assessment / Assessment	The material is praised, but guidance for feedback and peer assessment is considered insufficiently detailed.

The paraphrases in Table 6 summarize source texts and are not direct quotations. For R005, the Assessment–Technical disagreement is understandable because the review addresses both assessment design and broken links. For R116, A assigned neutral sentiment and B assigned positive sentiment, while both identified criticism and the Assessment aspect. These examples show why sentiment, aspect, and criticism should be retained as separate dimensions.

DISCUSSION

Limits of using ratings as proxies

The RQ1 results show that agreement between annotators in interpreting texts was higher than either annotator's agreement with rating categories in the same sample. This descriptive comparison does not establish that one representation is more accurate. Ratings express global evaluations, whereas reviews may evaluate several objects. The three-star cases show that the middle rating category cannot automatically be treated as text without positive or negative tendencies.

The corpus audit also distinguishes rating conflicts from proxy conflicts. Of the 402 groups with differing ratings, only 54 retained conflicting categories after mapping to three classes. This distinction prevents variation between four and five stars from being misread as conflicting sentiment. If the corpus is used to develop a classifier in future, all occurrences of identical texts should be placed in the same subset to prevent their distribution across training and test data. Rules for handling different labels should also be stated. This audit does not measure model performance.

Agreement and interpretive validity

The RQ2 results show high agreement in the available labels alongside uneven category distributions. Only six reviews according to A and seven according to B fell under Learning Support; Technical also contained only six and seven reviews, respectively. These counts limit

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

assessment of each category's stability. Following Sim and Wright (2005), κ should be interpreted with category marginals and case coverage rather than merely described as "high."

Agreement on criticism does not resolve its operational boundaries. R026, which describes the material as very theoretical and unstructured, was labeled Yes, while R023, which complains about complicated concepts and poor presentation, was labeled No. This difference between types of cases does not establish that the labels are incorrect, but it shows the need to explain how specific criticism must be to qualify. Further examination should distinguish general complaints, problems that can be located in course materials or processes, and explicit suggestions for improvement. Annotators should document this explanation; it is not assumed to be a protocol already used.

For aspect disagreements, selecting a single primary aspect can simplify reviews addressing multiple aspects. Multilabel coding or sentiment coding for each aspect could retain information lost when Technical and Assessment must be treated as mutually exclusive. Developing such categories is future work rather than new coding performed in this study. Discussion of disagreements should preserve annotators' interpretive reasoning rather than focus solely on increasing κ (O'Connor & Joffe, 2020).

Pedagogical implications and limits on action

The RQ3 results support three areas for examination. First, criticism in R003, R005, and R116 directs attention to alignment among objectives, content, quizzes, assignments, and peer assessment rubrics. Constructive alignment provides a basis for examining whether assessment demands are adequately supported by learning activities and materials (Biggs, 1996). This is a proposed review process; the review texts do not establish that course design is objectively misaligned.

Second, R001 and R036 direct attention to the pace of explanations, prerequisite material, explanations of notation, and opportunities for enrichment. R036 is relevant because its five-star rating and positive sentiment do not eliminate the need for specific improvements. Third, R066 and R084 direct attention to subtitles and forum facilitation. Feedback principles support specifying what needs improvement and how that improvement can be assessed (Hattie & Timperley, 2007).

These recommendations require confirmation against course materials, rubrics, and input from course instructors before implementation. The absence of course identifiers limits

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

assignment of actions to specific courses. Consistent with Jørnø and Gynther (2018), an issue label does not itself constitute an actionable insight without a context for action. The use of analytics also requires examination of instructors' actions and their effects (Wise & Jung, 2019). To test claims of improved decisions, a separate study could compare a dashboard with rating summaries using decision quality assessed against a rubric rather than user satisfaction alone.

LIMITATIONS

First, the selection mechanism for the annotation sample is unknown, and its stratum composition differs from that of the corpus. Criticism frequencies, discrepancy rates, and κ describe the 120 reviews analyzed; generalization of prevalence is not warranted. Second, the annotations do not constitute an adjudicated gold standard. Details of the original process, including separation of the assessments and access to ratings, were not measured, so agreement is not interpreted as proof of independence.

Third, a single primary aspect label simplifies reviews addressing multiple aspects, while the boundaries of criticism still require operational documentation and further examination. Perfect agreement does not remove these construct limitations. Fourth, the language check establishes only the presence of non-English examples; language proportions were not calculated. Repeated texts were retained because the purpose was to audit the corpus as recorded rather than identify unique learners.

Fifth, structured attributes describing courses, review dates, learner characteristics, and learning outcomes were unavailable. Incomplete metadata on provenance and the collection period limit external traceability and assessment of temporal relevance. Sixth, the study includes no evaluation of automated models, expert assessment of recommendations, implementation of interventions, or measurement of educators' decisions. The results therefore do not demonstrate dashboard effectiveness, improved learning, or the superiority of any algorithm.

CONCLUSION

In a sample of 120 reviews with equally sized rating strata, human sentiment judgments did not always agree with categories derived from ratings. The clearest differences occurred among three-star reviews, of which only 12 were labeled neutral by A and eight by B. Inter-annotator agreement reached κ = 0.9018 for sentiment, 0.8316 for primary aspect, and 1.0000

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

for the presence of criticism, but these values do not establish construct validity or process independence.

Sixty-six reviews contained criticism, including three with high ratings. Some texts labeled positive also contained criticism. This evidence supports presenting ratings, sentiment, aspects, and source texts together when examining learning experiences. Implications concerning assessment alignment, support, and access to materials provide traceable directions for examination, while the appropriateness of actions and improvements in decision quality require separate evaluation.

DATA AND ETHICS STATEMENT

The research supplement contains audit code, SHA-256 hashes identifying the input files, result summaries, and derived tables linking each Code to all matching source Id values. All original labels and Notes from Annotators A and B are retained. No new responses from learners or educators were collected for this analysis. The data have no respondent identity fields, but free text may still contain names or clues to identity; examples in the article are therefore presented as coded paraphrases. Access to and redistribution of the parent corpus are subject to the source data terms, which must be confirmed before public dissemination.

REFERENCES

Awadh, W. A., Sulaiman, R. B., & Mahmoud, M. A. (2025). Aspect-based sentiment analysis in MOOCs: A systematic literature review introducing the MASC-MEF framework. *Journal of King Saud University Computer and Information Sciences, 37*, Article 2. https://doi.org/10.1007/s44443-025-00018-1

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347–364. https://doi.org/10.1007/BF00138871

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Jørnø, R. L., & Gynther, K. (2018). What constitutes an “actionable insight” in learning analytics? *Journal of Learning Analytics, 5*(3), 198–221. https://doi.org/10.18608/jla.2018.53.13

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Li, L., Johnson, J., Aarhus, W., & Shah, D. (2022). Key factors in MOOC pedagogy based on NLP sentiment analysis of learner reviews: What makes a hit. *Computers & Education, 176*, Article 104354. https://doi.org/10.1016/j.compedu.2021.104354

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*. https://doi.org/10.1177/1609406919899220

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Wise, A. F., & Jung, Y. (2019). Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics, 6*(2), 53–69. https://doi.org/10.18608/jla.2019.62.4

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License