JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Technology Report

# Enhancing the Value of Large-Enrollment Course Evaluation Data Using Sentiment Analysis

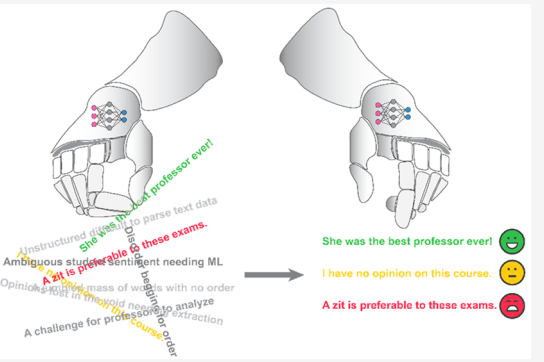
Benjamin B. Hoar, Roshini Ramachandran, Marc Levis-Fitzgerald, Erin M. Sparck, Ke Wu, and Chong Liu*

Cite This: *J. Chem. Educ.* 2023, 100, 4085−4091

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** In education, space exists for a tool that valorizes generic student course evaluation formats by organizing and recapitulating students' views on the pedagogical practices to which they are exposed. Often, student opinions about a course are gathered using a general comment section that does not solicit feedback concerning specific course components. Herein, we show a novel approach to summarizing and organizing students' opinions as a function of the language used in their course evaluations, specifically focusing on developing software that outputs actionable, specific feedback about course components in large-enrollment STEM contexts. Our approach augments existing course review formats, which rely heavily on unstructured text data, with a tool built from Python, LaTeX, and Google's Natural Language API. The result is quantitative, summative sentiment analysis reports that have general and component-specific sections, aiming to address some of the challenges faced by educators when teaching large physical science courses.




**KEYWORDS:** *First-Year Undergraduate, Professional Development, Administration Issues, Student-Centered Learning, Machine Learning*



## INTRODUCTION

Student course evaluations are a fundamental way in which students participate in their own education.[1,2] One limitation of this approach, however, is that students are often asked to provide feedback in an unstructured, open-ended manner.[3,4] This practice is limiting in that it does not allow faculty to gain quick insight into student opinions regarding specific components of their course such as lectures and homework. This issue is exacerbated in large-enrollment physical science courses at large research universities, where instructors may receive upward of a thousand course evaluations per year.[5,6−8] Further, larger classes contain students from diverse educational backgrounds, and certain classroom practices can be less supportive of underprivileged students, limiting their ability to succeed.[9] Compounding these factors is the growing reliance on virtual learning, which introduces another obstacle to instructors[10,11] and students,[12,13] especially students from economically disadvantaged groups.[14,15] Student course evaluations are a widely implemented tool available to instructors to gather summative data relating to the efficacy of their practices and iteratively adapt them.

While course evaluations do provide a possible means to improve teaching practices, concern over student biases regarding instructor race[16] and gender,[17] among other factors,[1] and the evidence that students' evaluation of their own learning can be poorly correlated with actual learning,[18] may decrease trust in student opinion. However, skepticism tends to focus on the use of numerical rankings and aggregate data for the use of institutional evaluation of instructors, with less seeing student evaluations as wholly uninformative.[1,19−21] Besides, text-based student feedback can be valuable for the inclusion of student voices from all socioeconomic and educational backgrounds in teaching practices.[14,20,21] If instructors are aware of these pitfalls and their own possible biases in reading evaluations,[22,23] then clear, interpretable reports on student sentiment ought to be sought as a valuable tool to improve course quality.

In light of these challenges, UCLA, whose enrollment (∼45,000) comprises students from diverse socioeconomic backgrounds,[24,25] is an appropriate setting for testing approaches to augmenting course evaluation practices with modern data tools. In UCLA physical science courses, text-based feedback is often open-ended with only a few supplemental numerical ranking questions presented, such as, "what is your overall rating of the [instructor] on a 0−9 scale". This results in an abundance of unstructured text data, challenging instructors to evolve their courses rationally, especially with the numerous biases and pedagogical obstacles already faced.

**Received:** March 29, 2023
**Revised:** August 31, 2023
**Published:** September 15, 2023

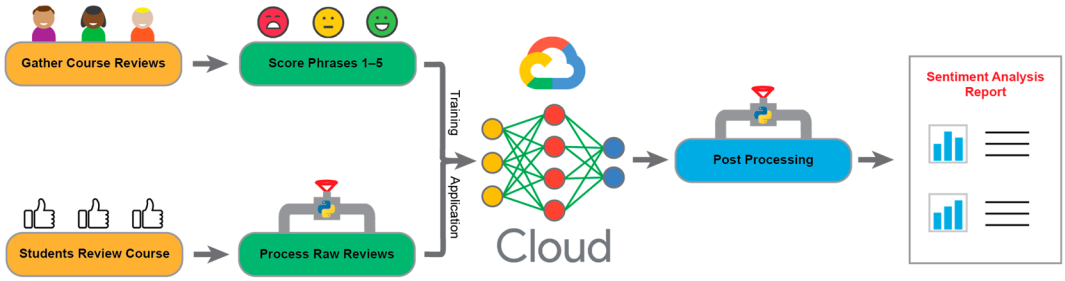



**Figure 1.** Overview of training and the general application pipeline. In the algorithm training phase, course reviews are gathered, scored, and used to train a GCSA algorithm which can then be subsequently used to generate a sentiment analysis report. After training, student course reviews can be processed and scored by the GCSA algorithm without human intervention.

Fortunately, software and machine learning provide the means to develop tools for augmenting existing course evaluation practices. Software provides custom organization and presentation of machine learning sentiment analysis data, data which ranks positivity of text data on a numerical scale: one (negative) to five (positive) herein.[26,27] In this work, the Google Cloud Platform's Sentiment Analysis API (GCSA) is used as the primary means of performing sentiment analysis. On demand, cloud-service algorithms are superior to other options in terms of flexibility and accessibility, permitting those with effectively no prior machine learning knowledge to develop powerful state-of-the-art machine learning models for implementation in their own custom software. Herein, the discussion will focus on the gathering of course evaluations for training data, the training of a GCSA model, and the output of custom Python scripts leveraging this data and service as a means to augment existing course evaluation practices while addressing some of the challenges facing educators in large-enrollment STEM contexts.

## BACKGROUND AND RELEVANCE

Machine learning is defined as "the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data".[28] In essence, machine learning algorithms can be trained on data where relationships are known (supervised machine learning) to predict a desired metric on arbitrary future data. A subset of machine learning, sentiment analysis, aims to learn how the vocabulary and language employed by an author correlates to positive, negative, or neutral opinion.[26,29,30] This technique is ubiquitous in companies aiming to gauge public opinion on their products or marketing strategies.[31] Often, these opinions are gathered from public platforms such as Twitter or Instagram and subsequently scored using sentiment analysis before presentation of summative sentiment in easy to interpret graphs and tables, which succinctly demonstrate how well a company's product, advertisements, etc., are being received by the public. Recently, sentiment analysis has also aided with measuring the efficacy of politicians' communication regarding COVID-19.[32] Finally, sentiment analysis has seen application in education contexts as well, having been used in the analysis of student feedback emotionality[33] and in classifying reactions to E-learning practices as binary positive or negative.[34] In addition to classifying sentiment as positive or negative, Kumar et al.[33] showed the evolution of student emotionality over time, focusing on feelings such as anger, fear, joy, and trust. While this was a valuable use of sentiment analysis, it does not address the issue of summarizing large amounts of text while capturing topic specific sentiment. Considering the challenges facing educators and the nascent implementation of machine learning tools in the context of course evaluation data analysis, a summative sentiment analysis machine learning tool is primed for implementation.

## METHODS

With the accessibility of GCSA in mind, the first step in implementing these techniques was to gather and score data representative of student course evaluations, initially with an emphasis on UCLA large-enrollment STEM courses.[35] By gathering and labeling a representative data set, a GCSA model could learn to predict student sentiment scores which ultimately enabled the production of sentiment analysis reports[36,37] (Figure 1).

In the case of this research, the input was student course reviews in the form of statements or sentences, and the output was the relative positivity of these statements on a scale of 1−5. To gather these data, nine UCLA professors and five teaching assistants were asked to provide student reviews of their courses from previous quarters. Initially, student course reviews were pulled from lower-divisional undergraduate courses in the physical sciences with class sizes on the order of one hundred to three hundred students. Professor data provided feedback on lecture and course structure, while teaching assistant evaluations focused on more interactive environments, such as lab and discussion sections (∼20−30 students). Data from these diverse teaching environments is likely to provide a more robustly trained machine learning algorithm.[38] Due to the generally positive nature of UCLA student course reviews, https://www.ratemyprofessors.com/ was also consulted to extract more negative and neutral reviews, around 15% of the total data set, to balance our training data; a sufficient representation of all possible sentiment scores is generally required to ensure model robustness.[39] In acquiring data from this public source, the primary emphasis was the collection of data that directly discussed course content and educator quality while remaining largely agnostic to the course subject and thus pulling data evaluating college courses in general. In the end, this resulted in a data set of 1,603 phrases (Figure S1). To minimize bias in the scoring of the data set (initial data collection and approximate scoring was performed by Benjamin Hoar), a total of 20 UCLA undergraduate students were gathered to perform additional scoring, a process that required approximately four and a half hours of scoring time per student assuming a rate of one statement scored per ten seconds. Directions on scoring stated that scoring should be conducted as objectively as possible (Supporting Methods, *scoring of training data*).[40] In the end, the median of the total of 20 scores (average standard deviation of 0.67) was counted as the *true* value of the phrase. After scoring,


https://doi.org/10.1021/acs.jchemed.3c00258
J. Chem. Educ. 2023, 100, 4085−4091

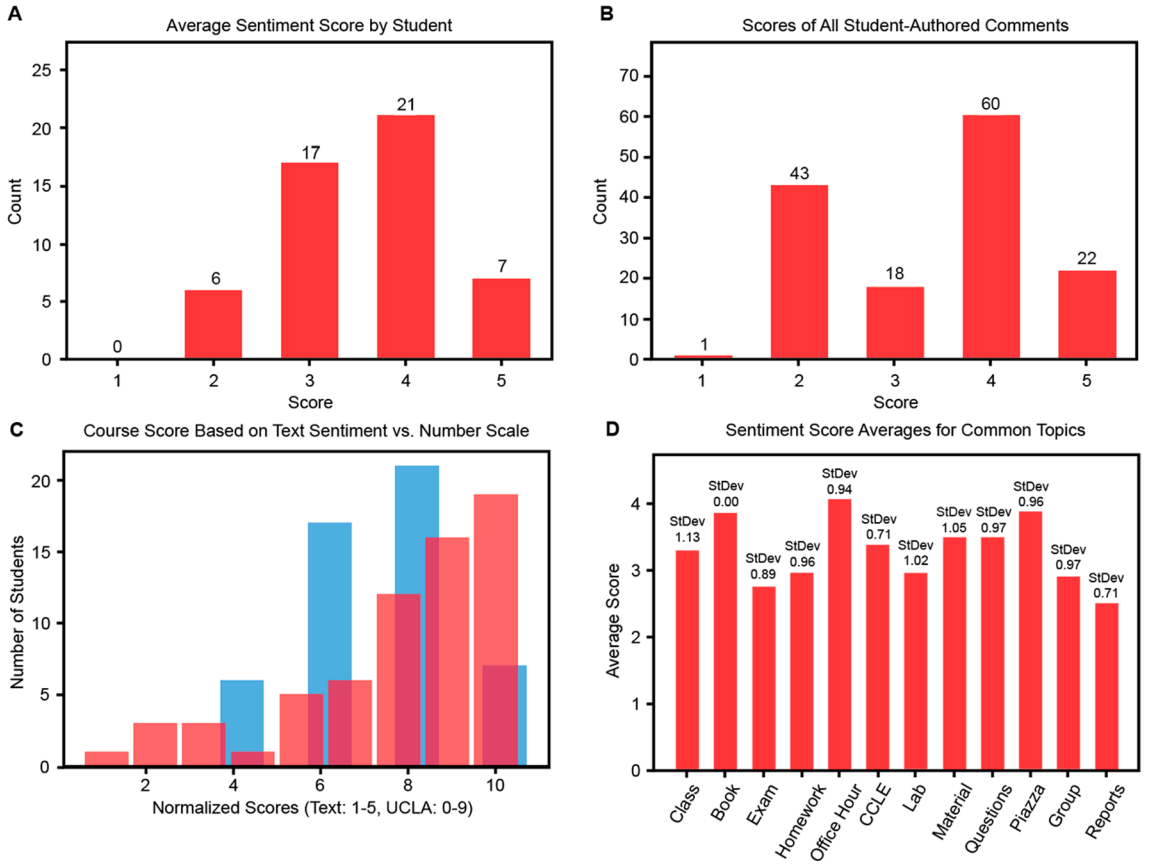



**Figure 2.** Overall course sentiment and average sentiment for selected course concepts. (A) Distribution of average sentiment of students {average: 3.57; std. dev.: 0.87}; in the figure, 17 students had a neutral average sentiment (score 3). (B) All scores of all phrases independent of author in text corpus {average: 3.41; std. dev.: 1.09}. (C) Comparison of student sentiment as a function of sentiment analysis performed on their written feedback (blue, same distribution as A) versus their numerical rating of the course on a 0−9 scale (red). (D) Average and standard deviation of student sentiment derived from phrases related to the terms shown along the *x*-axis.

this data was split into sets containing 1,282 training, 161 validation, and 160 testing instances. This split in data was necessary to select the best algorithm structure, with the training data used to extract correlations in the data and output, the validation set used to test the accuracy of the model *during* training, and the testing set used to test the accuracy of the model *after* training.[31] Using this split, the GCSA tool provided a trained sentiment analysis algorithm with an overall accuracy of 73.1% (Figure S2), in line with accuracies for modern fine-grained (i.e., nonbinary) sentiment analysis approaches.[27,41,42] It is worth noting that multiple scores contribute to positive (4 and 5) and negative (1 and 2) sentiment. This is important because a "true" label of "2" scored as a "1" is technically inaccurate but is not as egregious as a "2" receiving a score of "4," for example. While reducing the number of labels to three did increase accuracy to 82.5% (Figure S3), our discussions with educators convinced us to accept a lower accuracy to obtain higher granularity in scoring.

Following training, the sentiment analysis algorithm could be called from custom Python scripts developed specifically for student course review supplementation (Supporting Methods, *software development*). These Python scripts provided the means to convert any given raw text source into the format required for input into the scoring algorithm and subsequently its organization into a summative report (Supplementary Report) for instructor review. The first stage of this report generation involved the splitting of raw text data into phrases. This was accomplished by splitting on standard sentence-ending punctuation and the word "but". Following this, each statement was converted to a standard form and then scored by the GCSA algorithm and saved for subsequent report generation. The last step prior to report generation was selection of terms to display and calculation of summary statistics to be presented. Scored statements returned from the GCSA algorithm were filtered for terms that were of common interest (selected by the authors) and those that were of implied interest to each instructor via their prevalence in their course evaluations (selected by a word-cloud inspired algorithm). Once all scores were provided and filtered, they were organized into summative and component-specific figures and tables to provide instructors an overview of student sentiment.

## RESULTS

### General Results Section

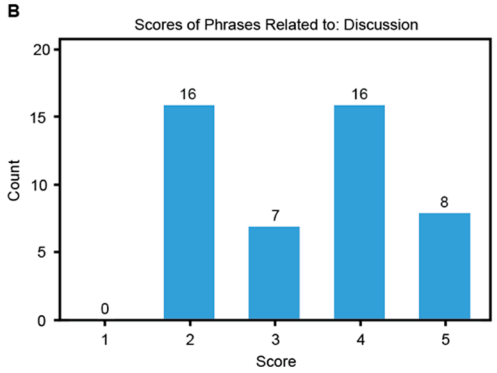
The "general" results section provided an overview of the entire text corpus and reflected student sentiment as a function of all student review statements. Here, four graphs are presented (Figure 2).

As an initial stage of report generation, all statements of all of the students were scored. These scored statements were visualized in two ways, normalized by author (Figure 2A) and independent of author (Figure 2B). The normalized version provided a visualization of student opinion, as implied by their


https://doi.org/10.1021/acs.jchemed.3c00258
J. Chem. Educ. 2023, 100, 4085−4091

A


B




C

| Comment | Score |
|---|---|
| He was an outstanding TA always willing to understand our situations and expand on course material. | 5 |
| He taught so efficiently and covered a lot of ground in a single discussion. | 4 |
| It seemed like we only went to discussion to work on a worksheet on our own. | 3 |
| Sometimes he would be a bit slow, worksheets would have helped. | 2 |
| He made a poor impression during discussion and his input was unhelpful and smug during lectures. | 1 |

**Figure 3.** Visualization of techniques applied to the component-specific section of the report. (A) A word cloud visually representing the selection of prominent terms for automated selection of course feedback terms that were not included in the predetermined list of words. (B) Representative distribution of scores related to the report component "discussion". (C) Representative table of phrases derived from student feedback that represent the commentary of students pertaining to the "discussion" component; an exemplar for score 1 was sourced from a different report than (B) for completeness.

individual average sentiments, while the alternative showed the prevalence of each sentiment score in the entire course evaluation corpus. While overall statement sentiment score distribution in Figure 2B was interesting, it could be biased due to highly verbose students, so must be considered as a complement to Figure 2A. In addition to their sentiment-based scores, students in UCLA physical science courses were asked "what is your overall rating of the [instructor] on a 0−9 scale?". This ordinal ranking did not provide any insight into the motivation behind rankings, and distributions of scores were typically shown to deviate from the sentiment analysis derived distributions (Figure 2C); however, text-based commentary was not necessarily limited to comments on the instructor. Finally, Figure 2D summarizes the component-specific results section. In Figure 2D, the average and standard deviation of scores related to a course component were presented as a prelude to the component-specific report section, which presented a more direct analysis of each of these terms. This panel was valuable in highlighting unanimously positive or negative components while also indicating possible sources of controversy due to high variability in student opinion, facilitating evaluation of the coming component specific section of the report. Figure 2, in general, provided the most concise overview of student sentiment possible. General student sentiment and component specific sentiment could be quickly analyzed, providing the instructor with a concise overview of how their students perceived the course and its components.

## Component Specific Results Section

In addition to this general section, component-specific reports were generated to provide insight into how specific course practices were received. Two methods of generating these reports were considered. First, from a discussion between UCLA faculty and advisors from the UCLA Center for the Advancement of Teaching, a list of globally relevant terms that encompassed course aspects common to most, if not all, university level physical science courses was established. The terms covered the instructor, lecture, textbook, exams, homework, CCLE (a course management portal), and office hours. In addition to these terms, six autoselected terms were also used to generate component-specific reports. These terms were selected via a word-cloud inspired algorithm (Figure 3A).[43] In a word cloud, the most common words in a body of text are presented in a graphic with the most common words appearing larger. To select for these terms, the counts of all words in the entire student-reviewed text corpus were calculated. From this list of word-counts, words already accounted for in the predefined list of valuable terms (e.g., homework) and meaningless "stop-words"[44] (words such as the, if, and, etc.) were also removed from consideration. What remained was a list of the most common course components not captured by the preselected list, which were implicitly considered to be of greatest interest to instructors. Common autoselected terms included "lab", "chemistry", "material", and "organic". Niche terms such as "Piazza" (a nonubiquitous student Q&A forum), "recorded" (referring to recorded lectures), and "clicker" (referring to a remote students used to answer questions live in class) were also selected for their appropriate instructors. These autoselected terms highlighted the value of this algorithm and utility of the tool in general, as they can capture both subject-level (e.g., "chemistry") and class-level (e.g., "Piazza") terms tailored to report recipients.

Considering these terms of interest, each component was given a page to demonstrate the overall sentiment of the students as it pertained to that component, here providing actual sample comments from students for the first time in the report. On each of these pages a score distribution (e.g., Figure 3B) was presented atop a table (Figure 3C) of exemplary phrases, providing samples from each of the five scores of the ranking system. The phrases selected for presentation were further fed through a pretrained, binary (positive or negative) sentiment analysis algorithm named VADER (Valence Aware Dictionary and sEntiment Reasoner),[45] which provided secondary assurance that the presented phrases in component specific report tables (e.g., Figure 3C) were accurate representations of their classes. VADER was only employed to filter out egregious GCSA misclassifications and was not trained on our data set or involved in any accuracy reporting; it is a well-known binary sentiment analyzer and was employed only as a quality control measure at the report generation stage. It is worth noting that VADER was an optional addition to our approach, and others may choose another way to filter the presented statements. Once phrases were selected, they offered the instructor insight into overall sentiment regarding a course component (Figure 3B) and showed specific statements that indicated strengths and weaknesses related to the execution of a course component (Figure 3C). Once all the data was processed, scored, and organized, the data was automatically converted into a


https://doi.org/10.1021/acs.jchemed.3c00258
J. Chem. Educ. 2023, 100, 4085−4091

standalone report with the aid of PyLaTeX,[46] a Python library that can be implemented to automate the creation of LaTeX documents. Within this document, directions on how to interpret the data and additional insight into report generation are also included.

### Effectiveness and Feedback

To gauge the utility of this proof-of-concept supplementary report, four UCLA professors (nine solicited) provided voluntary feedback on the reports generated from their course evaluations via a survey (Supporting Methods, *solicitation of feedback*). This survey was developed by the UCLA Center for the Advancement of Teaching to garner anonymous feedback about the efficacy of this tool. In total, we received four responses to the opt-in survey, which precluded any statistical analysis of the tool's reception but did provide insight into initial opinion (Table S1). In addition to their survey responses, two instructors provided longform commentary. Instructors unanimously agreed that the general format of this report was satisfactory, that the 5-point scale was sufficiently informative, and that the report was a valuable complement to their analysis of student feedback. While no aspect of the report was unanimously unsatisfactory, some limitations were noted. Respondents commented that the selected keywords and report were not exhaustive enough; both longform commenters wished to see a component report on "workload", and each provided additional terms such as "chemistry", "grading", and "collaboration" that they wished appeared in their reports. It is worth noting that, had those concepts been popular among student feedback, the autoselection algorithm would likely have captured them.

## FUTURE OUTLOOK

While we have established a useful proof of concept tool, there is room for advancement in our practices that would address the current limitations. For instance, custom user queries could be used to increase the flexibility of component specific reports. Additionally, parsed comments could be placed back in their original context, highlighted, and then presented to promote the contextual understanding of student sentiment.

The ideal path forward to improve adaptability and implementation of updates is to move this approach to a web-based format. This would expand user autonomy via the facile addition of user queries and perhaps even user accounts to track evaluation metrics over time. A web-based format would allow for the maintenance of favorable features and improvement of other features. In addition, increases in model accuracy could be obtained through the development of larger training sets with potentially minimized bias and the solicitation of machine learning experts to develop powerful, custom models specifically optimized for course evaluation sentiment analysis.

Alongside machine learning are necessary foundational improvements to course evaluation practices; improved solicitation of feedback will improve data quality and subsequently any machine learning tool trained on the data. UCLA is currently adopting more targeted close-ended and open-ended evaluation practices, which will be useful to obtain more informative feedback. Further, concrete, direct implementation of prompts regarding aspects such as inclusion and diversity is essential to obtaining valuable feedback regarding equity. While software can be used to extract sentences discussing homework (and its synonyms), for example, it is much more challenging to extract sentiment related to more abstract ideas that do not rely on simple, standardized language. Considering this, updates to the course evaluation formats so that students can elaborate on the achievement of learning outcomes, specific course practices, classroom climate, etc. would be the most effective approach for obtaining insight from diverse student populations while providing richer data to leverage with machine learning tools.[47]

## CONCLUSION

In short, a proof-of-concept tool has been developed to augment existing general course evaluation feedback data. By combining programming and machine learning principles, educators can now visualize how the language used by students in their course reviews relates to their feelings about course components in a dual quantitative and qualitative way. With this tool, educators can analyze graphical and tabular data about the opinions of their students with respect to the most common course components of interest. This format is especially valuable for large-enrollment STEM courses, as pedagogical approaches need to be tuned to broad audiences while maintaining effectiveness. Further, the autoselection of popular terms in feedback data allows for the presentation of course-specific reports that may not apply to every instructor but are invaluable to STEM instructors who employ a wide array of teaching aids to convey challenging subject matter. In recent times, rapidly changing demographics and teaching formats have provided an unprecedented challenge to educators who have faced diverse educational backgrounds, uncertainty related to the effectiveness of new teaching practices (both locally and remotely), and large numbers of students. Thankfully, machine learning approaches such as ours provide the opportunity for addressing these challenges, allowing for rapid, positive course evolution.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.3c00258.

Method details about the scoring of training data, software development, and solicitation of feedback; data set class populations and size; confusion matrixes; responses to ordinal ranking survey questions (PDF)

Supporting reference sample reports of student evaluation sentiment analysis (PDF)

## AUTHOR INFORMATION


### Corresponding Author

**Chong Liu** − *Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, California 90095, United States;* orcid.org/0000-0001-5546-3852; Email: chongliu@chem.ucla.edu

### Authors

**Benjamin B. Hoar** − *Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, California 90095, United States*

**Roshini Ramachandran** − *Center for the Advancement of Teaching, University of California Los Angeles, Los Angeles, California 90095, United States; Department of Biology and Chemistry, California State University, Monterey Bay, Seaside, California 93955, United States;* orcid.org/0000-0002-2559-4656




https://doi.org/10.1021/acs.jchemed.3c00258
J. Chem. Educ. 2023, 100, 4085−4091

**Marc Levis-Fitzgerald** − *Center for the Advancement of Teaching, University of California Los Angeles, Los Angeles, California 90095, United States*
**Erin M. Sparck** − *Center for the Advancement of Teaching, University of California Los Angeles, Los Angeles, California 90095, United States*
**Ke Wu** − *Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, California 90095, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.3c00258

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work is supported by UCLA Center for the Advancement of Teaching, Instructional Improvement Grants (19-11 and 19-11-01). C.L. acknowledges the support of the National Science Foundation (CHE-2247426) and the Sloan Research Fellowship from the Alfred P. Sloan Foundation.

## ■ REFERENCES

(1) Spooren, P.; Brockx, B.; Mortelmans, D. On the Validity of Student Evaluation of Teaching:The State of the Art. *Rev. Educ. Res.* **2013**, *83* (4), 598−642.
(2) Zabaleta, F. The use and misuse of student evaluations of teaching. *Teach. High. Educ.* **2007**, *12* (1), 55−76.
(3) Mirończuk, M. M. Information Extraction System for Transforming Unstructured Text Data in Fire Reports into Structured Forms: A Polish Case Study. *Fire Technol.* **2020**, *56* (2), 545−581.
(4) Li, Z.; Fan, Y.; Jiang, B.; Lei, T.; Liu, W. A survey on sentiment analysis and opinion mining for social multimedia. *Multimed. Tools Appl.* **2019**, *78* (6), 6939−6967.
(5) Personal website of Laurence Lavelle; https://lavelle.chem.ucla.edu/ (accessed August 2023).
(6) Toby, S. The relationship between class size and students, ratings of faculty: Or why some good teachers should not teach general chemistry. *J. Chem. Educ.* **1988**, *65* (9), 788−790.
(7) Toby, S. Class size and teaching evaluation: Or, the "general chemistry effect" revisited. *J. Chem. Educ.* **1993**, *70* (6), 465−466.
(8) Wang, L.; Calvano, L. Class size, student behaviors and educational outcomes. *Organ. Manag. J.* **2022**, *19* (4), 126−142.
(9) Clark, T. M. Narrowing Achievement Gaps in General Chemistry Courses with and without In-Class Active Learning. *J. Chem. Educ.* **2023**, *100* (4), 1494−1504.
(10) Kollalpitiya, K. Y.; Partigianoni, C. M.; Adsmond, D. A. The Role of Communication in the Success/Failure of Remote Learning of Chemistry during COVID-19. *J. Chem. Educ.* **2020**, *97* (9), 3386−3390.
(11) Schmidt, S.; Wright, Z. M.; Eckhart, K. E.; Starvaggi, F.; Vickery, W.; Wolf, M. E.; Pitts, M.; Warner, T.; Taofik, T.; Ng, M.; Colliver, C.; Sydlik, S. A. Hands-On Laboratory Experience Using Adhesives for Remote Learning of Polymer Chemistry. *J. Chem. Educ.* **2021**, *98* (10), 3153−3162.
(12) Garris, C. P.; Fleck, B. Student Evaluations of Transitioned-Online Courses During the COVID-19 Pandemic. *Scholarsh. Teach. Learn. Psychol.* **2022**, *8* (2), 119−139.
(13) Salas-Pilco, S. Z.; Yang, Y.; Zhang, Z. Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A systematic review. *Br. J. Educ. Technol.* **2022**, *53* (3), 593−619.
(14) Namboodiri, S. Zoom-ing Past "the New Normal"? Understanding Students' Engagement with Online Learning in Higher Education during the COVID-19 Pandemic. In *Re-imagining Educational Futures in Developing Countries: Lessons from Global Health Crises*; Mogaji, E., Jain, V., Maringe, F., Hinson, R. E., Eds.; Springer International Publishing: Cham, 2022; pp 139−158.
(15) Ramachandran, R.; Rodriguez, M. C. Student Perspectives on Remote Learning in a Large Organic Chemistry Lecture Course. *J. Chem. Educ.* **2020**, *97* (9), 2565−2572.
(16) McPherson, M. A.; Jewell, R. T.; Kim, M. What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes. *East Econ. J.* **2009**, *35* (1), 37−51.
(17) Boring, A. Gender bias in student evaluations in teaching. *J. Public Econ.* **2017**, *145*, 27−41.
(18) Deslauriers, L.; McCarty, L. S.; Miller, K.; Kestin, G. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (39), 19251−19257.
(19) Uttl, B.; White, C. A.; Gonzalez, D. W. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Stud. Educ. Evaluation* **2017**, *54*, 22−42.
(20) Marsh, C. J. A critical analysis of the use of formative assessment in schools. *Educ. Res. Policy Pract.* **2007**, *6* (1), 25−29.
(21) Benton, S. L.; Ryalls, K. R. *Challenging Misconceptions About Student Ratings of Instruction. 2016, IDEA Paper 58*; 2016.
(22) Humphreys, A.; Wang, R. J.-H. Automated Text Analysis for Consumer Research. *J. Consum. Res.* **2018**, *44* (6), 1274−1306.
(23) Kuang, S. Y.; Kamel-ElSayed, S.; Pitts, D. How to Receive Criticism: Theory and Practice from Cognitive and Cultural Approaches. *Med. Sci. Educ.* **2019**, *29* (4), 1109−1115.
(24) *Facts & Figures: An overview of the data that makes UCLA unique*; https://www.ucla.edu/about/facts-and-figures (accessed August 2023).
(25) Yang, J. Op-Ed: UCLA should address and combat student food insecurity. *Daily Bruin*; 2023; https://dailybruin.com/2023/01/30/op-ed-ucla-should-address-and-combat-student-food-insecurity (accessed August 2023).
(26) Serrano-Guerrero, J.; Olivas, J. A.; Romero, F. P.; Herrera-Viedma, E. Sentiment analysis: A review and comparative analysis of web services. *Inf. Sci.* **2015**, *311*, 18−38.
(27) Tang, F.; Fu, L.; Yao, B.; Xu, W. Aspect based fine-grained sentiment analysis for online reviews. *Inf. Sci.* **2019**, *488*, 190−204.
(28) *Oxford English Dictionary*; Oxford University Press; https://www.oed.com/search/dictionary/?scope=Entries&q=machine+learning (accessed August 2023).
(29) Qaisar, S. M. Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory. In *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2020; DOI: 10.1109/ICCIS49240.2020.9257657.
(30) Shaukat, Z.; Zulfiqar, A. A.; Xiao, C.; Azeem, M.; Mahmood, T. Sentiment analysis on IMDB using lexicon and neural networks. *SN Appl. Sci.* **2020**, *2* (2), 148.
(31) Géron, A. *Hands-On Machine Learning with Scikit-Lean, Keras & TensorFlow*, 2nd ed.; O'Reilly Media, 2019.
(32) Box-Steffensmeier, J. M.; Moses, L. Meaningful messaging: Sentiment in elite social media communication with the public on the COVID-19 pandemic. *Sci. Adv.* **2021**, *7* (29), No. eabg2898.
(33) Rani, S.; Kumar, P. A Sentiment Analysis System to Improve Teaching and Learning. *Computer* **2017**, *50* (5), 36−43.
(34) Kechaou, Z.; Ammar, M. B.; Alimi, A. M. Improving e-learning with sentiment analysis of users' opinions. In *2011 IEEE Global Engineering Education Conference (EDUCON)*, Amman, Jordan, 2011; DOI: 10.1109/EDUCON.2011.5773275.
(35) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349* (6245), 255−260.
(36) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **2018**, *140* (31), 9844−9853.
(37) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **2019**, *5* (1), 22.

(38) Tamkin, A.; Brundage, M.; Clark, J.; Ganguli, D. Understanding the Capabilities, Limitations, and Societal Impacts of Large Language Models. *arXiv* **2021**, arXiv:2102.02503.
(39) Chen, Z.; Duan, J.; Kang, L.; Qiu, G. Class-Imbalanced Deep Learning via a Class-Balanced Ensemble. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33* (10), 5626−5640.
(40) Denecke, K.; Deng, Y. Sentiment analysis in medical settings: New opportunities and challenges. *Artif. Intell. Med.* **2015**, *64* (1), 17−27.
(41) Angelidis, S.; Lapata, M. Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 17−31.
(42) Dang, N. C.; Moreno-Garcia, M. N.; De la Prieta, F. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics* **2020**, *9* (3), 483.
(43) Mueller, A. *WordCloud for Python*; GitHub, 2020. https://amueller.github.io/word_cloud/ (accessed August 2023).
(44) Kaur, J.; Buttar, P. K. A Systematic Review on Stopword Removal Algorithms. *Int. J. Future Revolut. Comput. Sci. Commun. Eng.* **2018**, *4* (4), 207−210 https://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1499.
(45) Hutto, C. J.; Gilbert, E. E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* **2014**, *8* (1), 216−225.
(46) Fennema, J. *PyLaTeX*, 1.3.2; GitHub, 2015; https://jeltef.github.io/PyLaTeX/current/ (accessed August 2023).
(47) *Student Experiences of Teaching Revision Project*; https://teaching.ucla.edu/student-experiences-of-teaching/ (accessed August 2023).


https://doi.org/10.1021/acs.jchemed.3c00258
*J. Chem. Educ.* 2023, 100, 4085−4091