Polarity in the Classroom: A Case Study Leveraging Peer Sentiment Toward Scalable Assessment

Zachariah J. Beasley, Les A. Piegl, and Paul Rosen

Abstract—Accurately grading open-ended assignments in large or massive open online courses (MOOCs) is non-trivial. Peer review is a promising solution but can be unreliable due to few reviewers and an unevaluated review form. To date, no work has 1) leveraged sentiment analysis in the peer-review process to inform or validate grades or 2) utilized aspect extraction to craft a review form from what students actually communicated. Our work utilizes, rather than discards, student data from review form comments to deliver better *information* to the instructor. In this work, we detail the process by which we create our domain-dependent lexicon and aspect-informed review form as well as our entire sentiment analysis algorithm which provides a fine-grained sentiment score from text alone. We end by analyzing validity and discussing conclusions from our corpus of over 6800 peer reviews from nine courses to understand the viability of sentiment in the classroom for increasing the information from and reliability of grading open-ended assignments in large courses.

Index Terms—Aspect extraction, crowdsourcing, educational data mining, massive open online course (MOOC), peer reviewing, sentiment analysis.

I. Introduction

Scalable grading of complex assignments (e.g., websites, essays, designs, open-ended questions) is notoriously hard in large, online academic environments [1]–[4]. Peer review is an often-cited solution, but can be unreliable due to a low number of reviewers and an unevaluated review form. Our work addresses both issues in a novel way by leveraging sentiment analysis in the peer-review process. By organizing our courses to include peer teaching and crowdsourced peer review (as in [5]), we obtain a crowd of reviewers large enough (30–40 on average, although we have found 20 to be sufficiently stable) to confidently assess a student's work and as a side effect stimulate learning in reviewing students [6]–[8].

Rather than apply natural language processing techniques to a work itself (e.g., an automated essay scorer like [9]), we apply the techniques to the meaningful content generated by a peer-reviewing crowd (Fig. 1). Our lexicon-based approach allows us to 1) collect sentiment to derive a fine-grained grade of peer-review text, 2) verify that grade with various metrics, and 3) suggest aspects important to the reviewers: candidates for addition to the review form. In this way, we can fulfill a "need for data collection tools that support teaching and learning" in engineering courses [10]. Our data-driven process addresses this need by providing the instructor with a better understanding of peer review and by increasing confidence in assigning a grade for complex assignments in large and massive open online courses (MOOCs).

The main contribution of this work is the description of a novel *process*: the development of a domain-specific lexicon and sentiment scorer to equip an instructor to build a review form through aspect extraction and to grade peer-review comments through sentiment analysis. It is not a one-size-fits-all review form or lexicon, but should be tailored to a specific course context. The lexicon-based approach was chosen primarily to improve the information with which numeric results can be justified.

II. Related Work

Because sentiment analysis is widely used to predict ratings from text in a large variety of domains, we attempt to briefly cover representative examples from social media, creative works (e.g., movies), and products. In an ideal world, we could simply take the best trained classifier from one of these categories and use it to determine student grades. However, there are a number of complications to this approach detailed below. When highlighting others' approaches, we report on either accuracy, precision, recall, or F1 score according to the evaluation prioritized in the author's own work.

A. Transfer Learning

Transfer learning in the field of machine learning is the process of applying an algorithm in a different domain than that in which it was trained. From an ethical standpoint, basing student grades on a classifier from a different domain could result in unpredictable and unfair results. From a practical standpoint, sentiment analysis is "highly sensitive to the domain from which the training data is extracted" [11, p. 31]. Others have found wide disagreement in results when applying the same publicly available sentiment analysis tools to different domains (e.g., accuracy of 80.4%, 62.5%, and 71.5% on movie reviews, tweets, and Amazon reviews, respectively) [12].

Still, various approaches to domain adaptation have been attempted. Most require labeled data from at least the source domain (e.g., [13] and [14] which link based on related features and related words, respectively). Others address this task when there is limited labeled data and note performance improvements over ignoring unlabeled data, but still do not address a situation such as ours with no labeled data [15]. Another alternative is a nearly unsupervised approach based on a two-phase clustering method, but a domain-dependent lexicon is still needed or else accuracy will "suffer a serious performance loss once the domain boundary is crossed" [16, p. 269]. A hybrid lexicon with weighted values from a local and
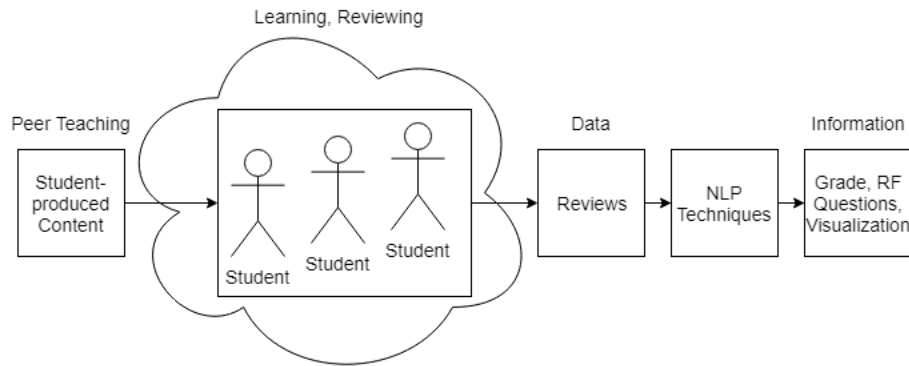
Fig. 1. Assessment process highlighting when natural language processing techniques are utilized.

general-purpose lexicon was utilized by [17], but performance improvements were found on only two out of three social media datasets (with a much lower F1 score on the third). In fact, over half of the combinations they tried decreased the F1 score, even though all lexicons were in the social media domain. Finally, there has been observed a "clear benefit to creating hand-ranked, fine-grained, multiple-part-of-speech dictionaries for lexicon-based sentiment analysis" [18, p. 282].

In their paper testing 24 algorithms on 18 datasets (including Twitter, Yelp, YouTube, Amazon, Digg, BBC, TED, and Myspace) to determine benchmarks in sentiment analysis, [19, p. 23, 27] note that "sentiment analysis methods cannot be used as 'off-the-shelf' methods, especially for novel datasets" and found that "methods are often better in the datasets [in which] they were originally evaluated" even for popular algorithms like SentiStrength [20], [21] and SO-CAL [18]. Others have found significant improvement with a crowd-labeled, domain-specific negative word lexicon compared to two other general-purpose lexicons classifying negativity in German media reports and party statements [22]. They, like us, strongly advocate a *process* and note that even "some commercial providers advise against using their sentiment lexicon out-of-the box without customizing it to the domain" [22]. Specifically, when narrowed down to sentiment analysis algorithms in the software engineering domain (Stack Overflow, Jira, etc.), [23] claim that publicly available sentiment analysis tools are inadequate and disagree with one another, but that domain-specific tools enhance accuracy. Others even go so far as to label the current state-of-the-art sentiment analysis tools deficient on the software engineering domain [24]. Taken together, this research shows the necessity of a domain-dependent lexicon and algorithm for accuracy in educational sentiment analysis.

B. Sentiment Analysis in Education

Influential and widely cited works summarizing sentiment analysis tools and techniques have often lacked any reference to its applications in academia [11], [25], [26]. However, in the past three to four years papers have begun to emerge leveraging sentiment analysis in academia, although to date it has not been applied to generate a grade based on review comment text or to provide any additional information to the

instructor. Similarly, we find no research leveraging aspect extraction to semiautomate review form creation in a data-driven way. The closest conceptual work to developing a review form from student input or data is a small-scale study in which 25 nursing students created a "a learner-driven feedback form" to reduce "the possibility of negative peer review experience" [27, p. 47]. In any case, it is important to improve both the quantity and quality of the information gathered from assessment data to allow instructors to pursue data-driven decision-making in the classroom.

Sentiment analysis has recently been used in the classroom to determine:

- student attrition over time in three MOOCs (captured from course forums and scored by a product reviews lexicon [28]) or predicted attrition in a single MOOC (captured from a course forum and scored by SentiWordNet 3.0 as one feature of a neural network [29]);
- the mood of students toward a teacher (captured via Twitter and scored by Naïve Bayes [30]) or as an "emotional thermometer for teaching" in virtual classrooms (captured in forum posts and scored by an ensemble [31]);
- negative students or course issues (captured in online course forums and scored by the Microsoft Text Analytics API [32] or captured from social media and scored by a mixed graph of terms [33]);
- teacher strengths and weaknesses identified by students (a proposed system with sentiment captured via questionnaire and scored by Naïve Bayes [34], or a proposed multilingual system with sentiment captured from Coursera peer reviews and scored by a lexicon in R [35], or a system with sentiment captured from teacher evaluations and scored by an ensemble [36]);
- student perception of internship experience (captured from transcribed interviews and scored manually [37]);
- an alternative way to view poetry and a means of student discussion on the relationship between text and numbers (captured from a Walt Whitman poem and scored by a proprietary sentiment analysis tool [38]).

There is one recent work which applies sentiment analysis to peer-review text, but it is limited to excluding highly negative reviews from being provided to the submitters so they will not disregard and disengage from the formative feedback

they receive [39]. It uses the VADER lexicon [40] to score content, which we have found inaccurate on our dataset when compared with other lexicons [41]. In addition, this work provides no other metrics or visualizations to the instructor and it does not contribute to a scalable, reliable grading process. While these applications of sentiment analysis in the classroom are interesting and helpful in their own ways, they do not contribute to an information gain from peer review toward improving the review form or providing a reliable grade. None go further than classical applications of sentiment analysis, merely noting subjective opinion toward an object.

III. METHODOLOGY

Our peer-review data was comprised of individual review form responses containing both analytical (radio button) and subjective (essay question) feedback. It came primarily from software engineering (SE), software testing (ST), computer graphics (CG), and geometric modeling (GM) courses at a large, R1 university computer science department over a period of five semesters. Of the nine courses, four were undergraduate only, the rest were cross listed for graduate and undergraduate students. Each course had approximately 40 students.

From these courses, we have a number of growing corpora of over 6800 peer reviews of 325 student works. The responses reflected peer's sentiment on three different types of projects: a weekly group presentation (35–40 reviewers), a semester-long group essay (20–40 reviewers), and a semester-long group term project (20–40 reviewers). Compared to other peer-review analyses, the number of reviews per student work was quite large [39], [42]. Textual responses were aggregated by our system to provide the mean, median, and standard deviation of sentiment, number of comments successfully scored, and various per-comment metrics (Section IV-D).

A. Review Form: Analytical Feedback

The analytical section of our review form was created using an iterative, data-driven approach [5]. Our seed growing algorithm began with a basic rubric and questions were added, modified, or removed after each semester through intelligent data combing: *a process of selecting information-rich keywords and phrases, through human intelligence, for the purpose of correctly analyzing and summarizing student observations*. This process was intentionally fuzzy—words were selected if they provided 1) meaningful sentiment (e.g., "extraordinary," but not "good") or 2) information (e.g., students mentioned a presence or lack of "citations" or "diagrams"). This was later semiautomated through the use of an aspect extractor detailed in Section IV-A, but still requires human intelligence for verification. The current review form has 22 questions divided into three categories: Overall score, Technical score, and Personalization score. Each question captures a student's response in an area that previous students have indicated is important [43].

B. Sentiment: Subjective Feedback

Selecting a review form radio button communicates a little information but allowing detailed feedback provides another

dimension of student response: sentiment. Though the analytic portion of the review form restricted students, the subjective section allowed freedom to discuss anything. Sentiment, along with a concise summary of what students actually said, provided rich information for an instructor to validate a peer-review score, particularly if the course was large enough that the instructor could not check every work. Thus, the sentiment score aggregated over the crowd's response was a component of our final score that could be used to validate or adjust the grade from the analytic section.

C. Assumptions

In our process, we aggregated document-level sentiment using an opinion lexicon (semantic orientation is according to instructor heuristic) for our domain. We had a regression problem on unlabeled data and presented an ensemble of scores to the instructor for reference, utilizing a soft voting scheme. We reported on regular (not comparative), direct (not indirect), and explicit (not implicit) opinions [11].

The final major assumption in our work was that we could not utilize an instructor grade as a ground truth for *every* individual's textual response. Many student reviewers, learning the material for the first time, presumably commented on fewer aspects of the work than the instructor, who had a better sense of the big picture. It was also possible that a student found something the instructor missed. Ultimately, we could not expect each student's review to conform to the instructor's. This prevented us from utilizing any grading system that measured reviewer competency (e.g., [42]), trained reviewers to review like an instructor, or penalized "poor" reviewers. The aggregate peer-review score was meant to accurately reflect the opinion of the crowd that consumed the information presented, although it could be adjusted by the instructor. Not using the instructor score as a ground truth also safeguarded against introducing bias into the algorithm. Our assumption of unlabled data also motivated the choice of a lexicon-based approach in addition to its intuitiveness, interpretability (since rationale for a grade was occasionally requested), and accuracy on short segments of text (see [44] for some other limitations of neural networks, especially identifying which linguistic properties are identified and explaining predictions). Ultimately, rather than try to use an existing recurrent neural network trained on a different labeled dataset (e.g., [45]), we chose an approach that would increase the amount of information available to the instructor.

IV. SENTIMENT ANALYSIS

A. Lexicon and Review Form

Over the course of five semesters, we gathered keywords for our lexicons through the process of intelligent data combing described above and weighted them by instructor heuristic [18], [46]. In contrast to the review form questions, which were selected for their breadth, we selected any words that exhibited positive or negative sentiment. However, we intentionally excluded overused words like "good" and "bad" that provide little quality information—[47] also found a balance of frequent and rare words necessary to discover subjective

content. Thus, our lexicons could be interpreted as a stemmed seed set which was not expanded through a lexical learning strategy since we desired a smaller set of words specific to our domain and weighted by heuristic (in contrast to [48]–[50]). Lemmatization was considered, but stemming was chosen for its ability to reflect more closely the actual written text. It allowed identification of keywords at a finer level of detail, including superlative adjectives (e.g., a "dry" vs. the "driest" lecture, which should be weighted differently). Similarly, the keywords "good," "better," and "best" all have the same lemma: "good." Thus, "better" and "best," which do exhibit significant sentiment, would have been excluded from analysis. Indeed, there would have been an average of 151 such cases per course in Spring 2019 of these two words alone being excluded if we had used lemmatization. Great care was taken when creating the lexicon to consider words in their variety of inflected forms.

Our positive word lexicon currently contains 250 words and our negative word lexicon currently contains 187 words. We have an additional lexicon comprised of words that negate sentiment (19 words) and a lexicon for flag words like "cheating" (12 words). The inclusion of flag words allows the potential for crowdsourced plagiarism detection [51]. When compared both qualitatively and quantitatively to six other lexicons publicly available (AFINN-111, ANEW-2017, MPQA, SentiWordNet 3.0, SlangSD, and Vader), our domain-specific lexicon provided more consistent tagging of high quality sentiment while appropriately ignoring neutral text [41], [52].

To semiautomate our process of intelligent data combing, we developed an aspect extractor similar to [53] that used a sliding window (Fig. 2) around sentiment-laden text (adjectives – italicized) to suggest aspects (nouns – highlighted) in close proximity for addition to the review form. This is a form of association rule mining between a noun and a set of sentiment words. There may be multiple—perhaps competing—sentiments on a single aspect. For example, in the review "...presentations were informative but dry...," both "informative" and "dry" are adjectives, with opposite polarity, associated with the noun "presentations." By finding the target of each sentiment word, rather than starting with the aspect, we were able to fine-tune the total sentiment per aspect. In Fig. 2, an adjective and noun match is found within the window. At the end of each semester, aspects were analyzed and if an aspect met a customizable threshold of mentions and absolute sentiment, it was considered a candidate for the review form and provided to the instructor (Table I). This process was also used to validate the current review form questions, which were all nouns, and to detect "parroting"—students simply mentioning keywords from the review form itself. Although bootstrapping the review form questions and lexicon by hand could not be avoided (and was in fact desirable for accuracy and intelligibility), the aspect extractor assisted with further iterations. Our aspect extractor focused on explicit aspects and ignored implicit aspects (in contrast to [50]). We evaluated our aspect extractor with the SemEval Aspect-Based Sentiment Analysis restaurant dataset, surpassing the highest precision and experiencing moderate recall in the new domain [54].

Fig. 2. Example sliding window over a review comment.

B. Lexicon Check

To classify the polarity of words, we utilized tokens (words and punctuation) from six categories (Fig. 3): positive sentiment, negative sentiment, neutral sentiment, negate word, flag word, and reset token. There was overlap between negative sentiment and negate words (e.g., "missing") as well as between negative sentiment and flag words (e.g., "copying"). Reset tokens were always neutral (e.g., "however"). We processed tokens in a specific order (Fig. 4) and saved tokens as either neutral, positive, or negative.
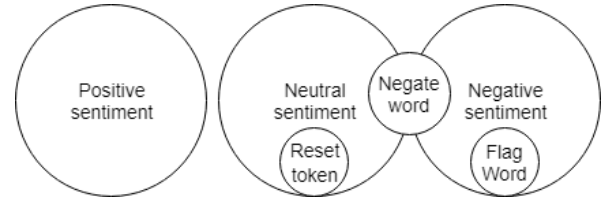

Fig. 3. Polarity of tokens.

C. Negation

Sentiment negating words, or "valence shifters," were perhaps the most complicated and interesting component of the sentiment grader and they occurred in just over 47% of our reviews. Words in our negate dictionary were comprised of regular negation words (e.g., "none"), presuppositional words (e.g., "barely"), omission words (e.g., "missing"), and modal auxiliary verbs (e.g., "should"). Negation is typically associated with negative sentiment, not distributed equally with positive sentiment [55], a finding we also observe in section VI. Similarly to [56], we found negating negative sentiment changed tone to neutral (e.g., "it wasn't terrible" or "not wordy") but negating positive sentiment changed tone to negative (e.g., "it wasn't clear" or "nothing innovative"). Our process captured negated positive words in three ways (see [18] for alternatives):

- negate word to reset token;
- preceding negative qualifier;
- trailing negative qualifier.

Certain words negated sentiment until their meaning was removed with a reset token (e.g., [57]):

> "...could have given ***more** practical examples to make it more clear as several readings were required to understand the topic and their idea flow*."

The positive words/phrases: practical examples, clear, understand the topic, and idea flow were all negated until the reset token, '.', was encountered. Other reset tokens that conclude a negate word's effect include: '.', ';', but, although, however, and nevertheless. Negate words include: no, not,

TABLE I
Example Aspects from an Entire Semester

| Noun	Occurrences	Context
presentation	164	["lucid", 0.7, "JJ", "the presentation is lucid and provided examples"]
essay	88	["balanced", 0.9, "JJ", "was a very balanced essay."]
topic	60	["relevant", 0.5, "JJ", "clearly explained the relevant topic briefly."]
team	54	["successful", 0.9, "JJ", "team has been successful in capturing all"]
animation	10	["impressive", 1.0, "JJ", "it was really impressive and the animation"]
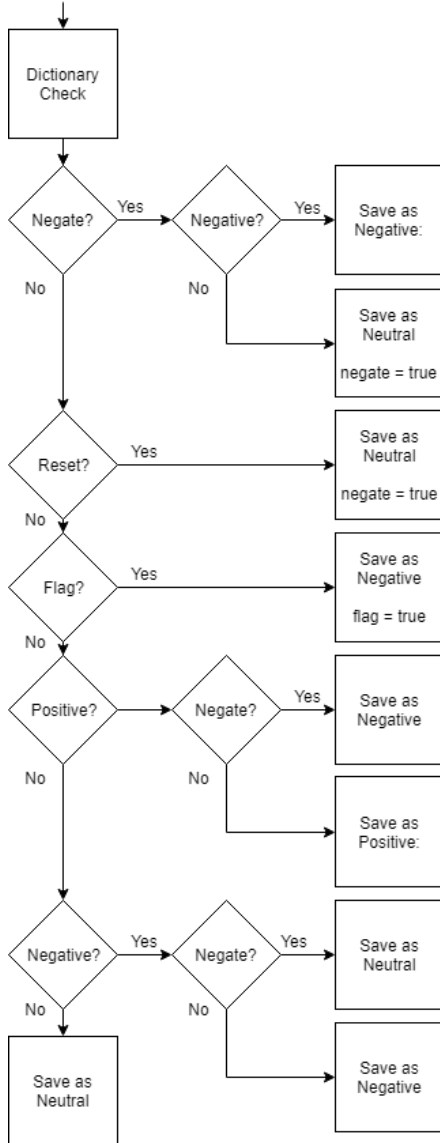
Fig. 4. Flowchart for the lexicon check.

can't, nothing, hardly, barely, lack, more, suggest, miss, and few.

Secondly, polarity was negated through a preceding negative qualifier. Simply put: a negative adjective in close proximity before a positive word ("close" is intentionally vague and is configurable in the algorithm):

"...which makes it **difficult** to understand..."

Finally, positive tone was negated through a trailing negative qualifier. This required observation into the future as the algorithm scanned the sliding window of text for a negative adjective which negated positive sentiment:

"...some *insight was* **missing**..."

It is important to note that the preceding/trailing negative qualifier was itself a negative sentiment word, thus "stacking" the effect of its negativity. This was intentional, as adjectives are often sentiment magnifiers [58], [59]. In contrast, negate words could either be negative or neutral, as seen in the overlap of the negate word lexicon in Fig. 3.

D. Grading

Our grading process (Fig. 5) extracted the analytical and subjective content into two files, which were processed by two different graders. The analytic grader simply matched student responses with their assigned values and aggregated the score. The process was relatively straightforward and is detailed in a prior paper [5]. Ultimately, the sentiment score (with accompanying metrics) was weighted with the analytic score and provided to the instructor as a suggested final grade from the peer-review process.

We utilized a variety of metrics and attributes to provide the maximum information to an instructor. Some are widely used in sentiment analysis (e.g., 'tone' and 'purity'), while others were specific to our process (e.g., 'dif'). The following metrics contributed to a sentiment score *per review comment*:

- weight: the sentiment of a lexicon-matched keyword: [0, 1];
- pos_keywords: the number of positive keywords: 0 to inf;
- neg_keywords: the number of negative keywords: 0 to inf;
- keywords: the total number of keywords matched: 0 to inf;
- tone: the sum of all weighted keywords: (-inf, inf);
- info: the absolute value of all weighted keywords: [0, inf);
- score: the sentiment score (F to A+): [0, 4.3].

The following attributes were defined or derived for validation and analysis *per review comment*:

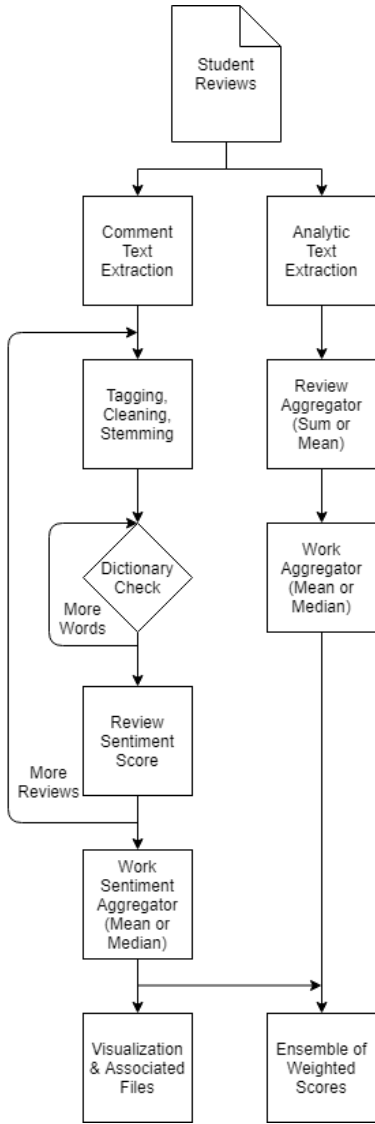- reliable: whether a comment has enough information to be aggregated: 0 or 1;

Fig. 5. Flowchart for the grading process.

- default: whether a comment matched a threshold of keywords and was scored: 0 or 1;
- dif: the difference between sentiment score and analytic section score: [-4.3, 4.3];
- purity: tone over info, a measure of how consistent the sentiment is: [-1, 1];
- positivity: the sum of positive sentiment: [0, inf);
- negativity: the sum of negative sentiment: (-inf, 0];
- negate_words: the number of negating words: 0 to inf;
- words_per_sentence: the number of words per sentence: 0 to inf;
- length: the number of words in the review: 0 to inf.

The following attributes were defined for validation and analysis *for an aggregation of peer reviews*:

- pos_dict_used: the percentage of tokens in the positive lexicon utilized: [0, 1];
- neg_dict_used: the percentage of tokens in the negative lexicon utilized: [0, 1].

Most importantly, our lexicon check for a single review

comment provided *tone* and *keywords*. We assigned a fine-grained score to the text (Fig. 6) by *tone over keywords* with some adjustments including scaling to the range desired by the instructor. Thus the highest scores were a result of many highly weighted positive keywords (i.e., highly positive purity), and lowest scores were those with many highly negative keywords (i.e., highly negative purity). We tested two options for aggregating comments by median and mean:

1) weight reviews based on information available (complex);
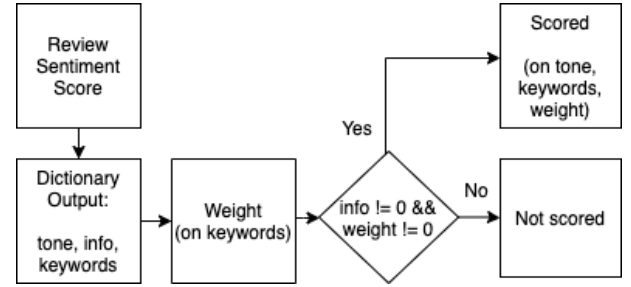2) weight all reviews equally (simple).

Fig. 6. Flowchart for the sentiment scoring algorithm.

In the first grading scheme, we began by weighting our confidence in the comment. If the sentiment was negative and lacked many keywords, we reduced the weight significantly to avoid penalizing harshly. If the review was positive with little information, we weighted slightly less to compensate. The second grading scheme in effect tested the principle of the wisdom of the crowd and allowed all reviews, even ones with a single keyword, to contribute to a student's score. This required a leap of faith—trusting that a few reviews with highly negative sentiment would not destroy a student's grade. In either scheme, if there was not enough information to process the comment we simply incremented the number of default scores. Over the last five semesters, roughly 75% of our reviews meet our basic threshold for scoring with weighted confidence. This number increased to approximately 85% when we incentivized students to provide a high-quality comment with a completion point. Reviews were checked by the teaching assistant and students quickly adjusted to providing in-depth reviews. Fig. 7 is an example of the sentiment grader on a single student's review (net positive – blue and underlined, net negative – red, negated – italicized), with the score and reliability generated from the complex scorer.

When we simplified to the second grading scheme for our two most recent courses where students were incentivized to provide quality review comments, the means decreased on average, although the difference was minute: $\sim$0.053 (1.2%) for Software Engineering (SE), an undergraduate course, and $\sim$0.068 (1.6%) for Geometric Modeling (GM), a graduate course. The medians were also affected, although less notably (0.68% and 0.15%, respectively). Fig. 8a and Fig. 8b show the difference in means using both scorers. One reason we saw some jumps between the grading schemes was that reviews with fewer keywords (especially negative reviews) were previously weighted significantly lower (or even disregarded)

"[The team] demonstrated that they knew the material well. I noticed that the presentation started out very strong and interesting until the group changed presenters and it became unengaging. I do have to point out that a lot of the images were informative but they were *not very pleasing* to look at and *not easy to understand*. Maybe some *more colorful images* would have *helped*. Other than this, the group did a good job at describing global and local approximation. In addition, their design examples were very good and interesting as well as their lab."

Reliable: 1
Grade: 3.36 / 4.30 (B+)

Fig. 7. Example of the sentiment scoring algorithm applied to a student's review. Net positive is blue and underlined, net negative is red, and negated sentiment is italicized.

so they would not sway the final score. A side effect of this simpler grading scheme was that our average percentage of reliable reviews went down along with our positive sentiment and average words because we kept terser, more negative reviews.

Simplifying the grader also had the side effect of increasing the average standard deviation of SE by 2.00% to 0.435 and GM by 3.07% to 0.479. Although the average standard deviation was slightly high and above our target of 10% of the grade range, or 0.430, we relied on the principle of the wisdom of the crowd to postulate that the means were still accurate. The fact that major algorithm changes (especially simplifications) only shifted grades within a fraction of a letter grade ($< 2\%$) suggested that we had a robust sentiment analysis algorithm when coupled with a large number of peer reviews per work.

V. VALIDATION

Firstly, sentiment score roughly tracked the review form score. In SE and GM, the review form score was slightly higher on average than sentiment mean (0.323 and 0.121 in the weighted versions and 0.376 and 0.190 in the unweighted versions). This ranged from 2.82% to 8.73% of total score. We did note some deviation between the *individual* scores—a phenomenon also observed in other domains [60], [61]. This deviation suggested that students were willing to write honestly, even if they gave their peers a good grade via the analytical section. Unfortunately, this fact prevented us from using the review form analytical score as a ground truth against the sentiment score. It also precluded us from using an algorithm from another domain, since we would have no way of quantitatively verifying the sentiment score.

However, in a prior work [41], we quantitatively compared the mean average error (MAE) between the *aggregate* review form analytic score and *aggregate* review form sentiment score for six publicly available lexicons: Affective Norms for English Words (ANEW) [62], SlangSD [63], Multi-Perspective Question Answering (MPQA 3.0) [64], Valence Aware Dictionary and sEntiment Reasoner (Vader) [40], SentiWordNet 3.0 [65], and AFINN-111 [46] while holding the scoring algorithm
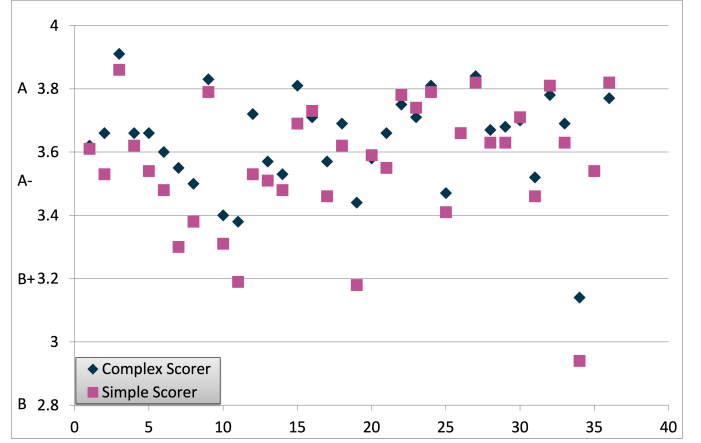
(a) GM (average difference -0.068)
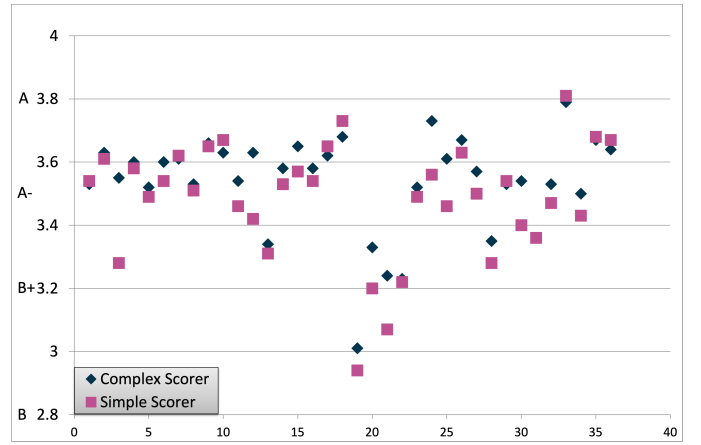
(b) SE (average difference -0.053)

Fig. 8. Change in (a) Geometric Modeling and (b) Software Engineering mean project scores (y-axis) using different scoring algorithms for each review (x-axis).

constant. This work showed that our lexicon captured roughly the same *amount* of sentiment even though the other lexicons were much larger [41]. GM and SE reviews carried enough information to be counted. We average 4.24 positive keywords, 1.48 negative keywords, and 1.13 negate words per review. Percent default scores ranged from 0% to 13% (3.6% on average for CG and 2.1% for SE) with 63% of reviews marked as reliable for the simplified scorer.

Additionally, the *quality* of the sentiment collected was the most precise from our lexicon. Fig. 9 is one qualitative example that demonstrates the difference in how the top three sentiment-producing lexicons—ours, ANEW, and SentiWordNet—interpreted text from the same review. This demonstrates each lexicon's capability of "understanding" the text by correctly noting sentiment. The closer the interpretation of positive and negative sentiment matches human interpretation, the better the lexicon. SentiWordNet captured the most sentiment, but not in an intuitive way, labeling words like "essay," "eye-catching," and "down" as negative or "due" and "use" as positive. ANEW, with keywords selected and weighted by hand, matched words more intuitively, but did

not capture clear sentiment-bearing words like "informative" or "nicely." ANEW also coded "team" as positive rather than neutral. In contrast, our lexicon correctly identified a large quantity of keywords and appropriately tagged the quality of sentiment while ignoring neutral words. It captured both the specific examples provided by the team and the lack of references as well as the positive keywords "excellent" and "informative."

Finally, the infrequency of instructor adjustment to the algorithm score was evidence for its correctness. In five semesters of use (nine courses, 324 student submissions), the instructor only changed the algorithm's score *once*. In this situation, the score was adjusted up to compensate for a number of students reporting a failing score for "reading off the slides". This problem was further compounded by the rather coarse-grained overall score selection (scale of four instead of the current eight) from a previous review form iteration. Increasing the granularity of overall score softened the otherwise large variation. Ultimately, testing different grading methods with our lexicon and finding they produce scores that are almost identical is an indication of a statistically sound procedure [66]).

VI. DISCUSSION

We evaluated the Pearson correlation coefficient on each metric from SE and GM with the mean and median simplified sentiment scores and used a significance of correlation test (degrees of freedom $(df) = 70$ and critical value of $+/-$ .380). Table II shows a selection of correlation values (r) that are statistically significant at the $\alpha = .001$ level. Mean and median scores correlated very closely (.935), suggesting that either may be utilized as the final score and that we have an approximately normal distribution. Purity was also highly correlated with mean which provided our first indication that we could simplify the scoring algorithm.

To discover what information is added from the simple sentiment scorer, we analyzed the correlation of mean with various metrics to infer the following:

1) Negate keywords have a similar negative correlation with the mean compared to Negative Words. This suggests that the role of negating words factors highly into grading, and situations in which they occur should be carefully handled;

2) Standard Deviation is negatively correlated with the mean. This suggests that most reviews are positive and that increased deviation comes from additional, negative reviews;

3) Percent Reliable has a very slight negative correlation with the mean (-.127). This suggests that most reliable reviews are negative and most unreliable reviews are positive;

4) Finally, a number of length and keyword metrics are negatively correlated with the mean (significant at the $\alpha = .05$ critical value of $+/-$ .232): Total Keywords (-.249), Words (-.284), Words/Sentence (-.258), and especially Adverbs (-.373). This suggests that *the more a student writes, the more faults they find in others' works* (even

though just 15% of all reviews contain more negative sentiment than positive sentiment).

Taken together, this information indicates that many students write cursory, unreliable positive reviews, but a few diligent students write longer, more detailed, and more negative reviews. We find such negative reviews more helpful in an academic context, and would be interested in analyzing whether those reviewers grade more like an instructor or are higher performing students. These reviews can be filtered to present to the instructor a more balanced and realistic perspective of the students' works.

We also analyzed the top one, five, ten, fifteen, and twenty percent of the most positive and most negative reviews to determine the top three keywords (including ties) most represented (Table III). GM had 920 reviews, while SE had 1027. It should be noted that only positive or negative keywords are included, with negated text excluded. For example, the following student's review is negative but only contains negated positive sentiment and not negative keywords: "They took a safe road quite literally. The safe itself looks standard and nothing like high end or unique but ifs [sic] a safe. They could have added more features to make it unique."

On the negative side, "hard," "need," and "short" quickly saturated the most negative reviews in both courses. On the positive end of the spectrum, "impressive" and "unique" seemed important to the reviewers in GM (who were designing models), while SE students (who were creating software systems) highlighted "useful" work that was easy to "understand." Students in both courses appreciated the inclusion of "examples." This type of insight can be utilized in courses even where a different peer-review format is utilized. In fact, our process, specifically aspect extraction and textual metrics, was utilized to deliver insights in three Data Visualization courses on factors related to student engagement in and enjoyment of peer review [67].

VII. CONCLUSION

Sentiment analysis is uniquely suited to increase and summarize the information gathered from peer-review text in large courses, which allows greater confidence in accurately grading open-ended assignments. Applying natural language processing techniques to the data generated from an intelligent, reviewing crowd increases the ability to capture content less accessible to automated graders: humor, beauty, sarcasm, or truthfulness. A lexicon-based approach increases information and allows for aspect extraction, which semiautomates review form modification based on features in actual student reviews. By utilizing our generalizable *process*, an instructor can approach assessment through crowdsourced peer review in a data-driven way. Although our approach was created in the context of a computer science course, mining information from peer-review text can yield a system tuned to any course's context. Ultimately, extracting sentiment from peer-review text is an untapped area with the potential to increase information and reliability in assessment.

In the future, we would like to better utilize the aspect generator via double propagation to use 1) known aspects to

OurLexicon

Grade: 3.49/4.00 (B+)

The team provided a condensed essay without any references. The graphics while very eye-catching, were not referenced in the text so I am unsure where they came from. However, the team provided an informative essay that broke down scripting in debugging nicely. In their 'Benefits of using script in debugging' section, the team provided the upsides of using automated testing in debugging by providing examples that included the fact that since automated scripts did not require human interaction to complete that they could be run overnight to reduce the cost of supervision and to optimize the time one had to run tests. Their 'Application example' section provided an excellent resource for an automated testing application in Selenium due to its use across multiple languages.

SentiWordNet

Grade: 3.08/4.00 (B)

The team provided a condensed essay without any references. The graphics while very eye-catching, were not referenced in the text so I am unsure where they came from. However, the team provided an informative essay that broke down scripting in debugging nicely. In their 'Benefits of using script in debugging' section, the team provided the upsides of using automated testing in debugging by providing examples that included the fact that since automated scripts did not require human interaction to complete that they could be run overnight to reduce the cost of supervision and to optimize the time one had to run tests. Their 'Application example' section provided an excellent resource for an automated testing application in Selenium due to its use across multiple languages.

ANEW

Grade: 3.53/4.00 (A-)

The team provided a condensed essay without any references. The graphics while very eye-catching, were not referenced in the text so I am unsure where they came from. However, the team provided an informative essay that broke down scripting in debugging nicely. In their 'Benefits of using script in debugging' section, the team provided the upsides of using automated testing in debugging by providing examples that included the fact that since automated scripts did not require human interaction to complete that they could be run overnight to reduce the cost of supervision and to optimize the time one had to run tests. Their 'Application example' section provided an excellent resource for an automated testing application in Selenium due to its use across multiple languages.

Fig. 9. Qualitative comparison of three lexicons. Net positive is blue and underlined, net negative is red.

TABLE II
Statistically Significant Correlations with Mean and Median

	FormScore	StdDev	NegSenti	Senti	Purity	NegKey	NegateKey	Adverbs
Mean	.750	-.462	.830	.764	.982	-.814	-.829	-.373
Median	.696	-.281	.820	.696	.913	-.821	-.813	-.432

find other aspects, 2) known aspects to find other sentiment words, and 3) known sentiment words to find other sentiment words via connectors (e.g., "and", "but", "however"). We are also interested in adopting a visualization to condense all peer reviews into an understandable graphic (perhaps using aspect-level opinion mining) to increase the information presented to an instructor for grading confidence.

VIII. ACKNOWLEDGMENT

REFERENCES

[1] N. B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, and M. J. Wainwright, "Some scaling laws for MOOC assessments," in

TABLE III
Keywords in the Most Positive and Negative Reviews

	Most Positive	Most Negative
GM 1%	unique(3), creative(3), outstand(3)	need(1)
GM 5%	unique(15), outstand(11), creative(9)	hard(3), need(3), short(3)
GM 10%	unique(25), creative(19), structure(18), example(18)	short(14), hard(7), need(5)
GM 15%	example(37), unique(37), impressive(32)	hard(16), need(15), short(15)
GM 20%	impressive(60), example(57), unique(54)	hard(21), need(21), short(21)
SE 1%	fantastic(3), engage(2), enjoy(2), useful(2), outstand(2)	punctuation(1), incorrect(1), awkward(1), heavy(1), bor(1), stuck(1)
SE 5%	useful(15), outstand(10), enjoy(9), unique(9), fantastic(9)	need(10), short(6), add(3), bor(3), heavy(3), hard(3)
SE 10%	useful(25), example(25), enjoy(18), outstand(18)	need(21), short(16), hard(11)
SE 15%	example(52), useful(32), outstand(30)	need(36), short(24), hard(15)
SE 20%	example(87), understand(40), useful(40)	need(48), short(30), hard(23)

KDD Workshop Data Mining for Educational Assessment and Feedback (ASSESS '14), New York City, NY, USA, 2014.

[2] G. Singh, S. Srikant, and V. Aggarwal, "Question independent grading using machine learning: The case of computer program grading," in *Proc. 22nd ACM Int. Conf. Knowledge Discovery and Data Mining (SIGKDD '16)*, San Francisco, CA, USA, Aug. 13, 2016, pp. 263–272, doi: 10.1145/2939672.2939696.

[3] C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer, "Scaling short-answer grading by combining peer assessment with algorithmic scoring," in *Proc. 1st ACM Conf. Learning at Scale (L@S '14)*, Atlanta, GA, USA, Mar. 4, 2014, pp. 99–108, doi: 10.1145/2556325.2566238.

[4] Z. J. Beasley, L. A. Piegl, and P. Rosen, "Ten challenges in cad cyber education," *Comput.-Aided Des. and Appl.*, vol. 15, no. 3, pp. 432–442, May 2018, doi: 10.1080/16864360.2017.1397893.

[5] ——, "Board 39: Designing intelligent review forms for peer assessment: A data-driven approach," in *Proc. 2019 ASEE Annu. Conf. & Expo.*, Tampa, FL, USA, Jun. 15, 2019, doi: 10.18260/1-2--32337.

[6] J. Russell, S. Van Horne, A. Ward, E. Bettis III, and J. Gikonyo, "Variability in students' evaluating processes in peer assessment with calibrated peer review," *J. Comput. Assisted Learn.*, vol. 33, no. 2, pp. 178–190, Apr. 2017, doi: 10.1111/jcal.12176.

[7] O. Almatrafi and A. Johri, "Systematic review of discussion forums in massive open online courses (MOOCs)," *IEEE Trans. Learn. Technol.*, Jul. 2018.

[8] H. Li, Y. Xiong, C. V. Hunter, X. Guo, and R. Tywoniw, "Does peer assessment promote student learning? a meta-analysis," *Assessment & Eval. Higher Educ.*, pp. 1–19, Feb. 2019, doi: 10.1080/02602938.2019.1620679.

[9] S. Somasundaran, M. Flor, M. Chodorow, H. Molloy, B. Gyawali, and L. McCulla, "Towards evaluating narrative quality in student writing," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 91–106, Jan. 2018, doi: 10.1162/tacl_a_00007.

[10] M. M. Moyne, M. Herman, K. Z. Gajos, C. J. Walsh, and D. P. Holland, "The development and evaluation of DEFT, a web-based tool for engineering design education," *IEEE Trans. Learn. Technol.*, vol. 11, no. 4, pp. 545–550, Feb. 2018, doi: 10.1109/TLT.2018.2810197.

[11] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Human Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[12] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Inf. Sci.*, vol. 311, pp. 18–38, Aug. 2015, doi: 10.1016/j.ins.2015.03.040.

[13] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proc. 45th Annu. Meeting Association Computational Linguistics*, Prague, Czech Republic, Jun. 2007, pp. 440–447.

[14] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proc. 19th Int. Conf. World Wide Web (WWW '10)*, Wanchai, Hong Kong, Apr. 26, 2010, pp. 751–760.

[15] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in *Proc. 1st Workshop Graph Based Methods for Natural Language Processing*, Stroudsburg, PA, USA, Jun. 2006, pp. 45–52, doi: 10.3115/1654758.1654769.

[16] A. Mudinas, D. Zhang, and M. Levene, "Bootstrap domain-specific sentiment classifiers from unlabeled corpora," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 269–285, Jan. 2018, doi: 10.1162/tacl_a_00020.

[17] A. Muhammad, N. Wiratunga, and R. Lothian, "Contextual sentiment analysis for social media genres," *Knowl.-Based Syst.*, vol. 108, pp. 92–101, Sep. 2016, doi: 10.1016/j.knosys.2016.05.032.

[18] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comp. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011, doi: 10.1162/COLI_a_00049.

[19] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto, "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Sci.*, vol. 5, no. 1, p. 23, Dec. 2016, doi: 10.1140/epjds/s13688-016-0085-1.

[20] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. Amer. Soc. Inf. Sci. & Technol.*, vol. 63, no. 1, pp. 163–173, Jan. 2012, doi: 10.1002/asi.21662.

[21] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. & Technol.*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010, doi: 10.1002/asi.21416.

[22] M. Haselmayer and M. Jenny, "Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding," *Qual. & Quantity*, vol. 51, no. 6, pp. 2623–2646, Nov. 2017, doi: 10.1007/s11135-016-0412-4.

[23] N. Novielli, D. Girardi, and F. Lanubile, "A benchmark study on sentiment analysis for software engineering research," in *IEEE/ACM 15th Int. Conf. on Mining Software Repositories (MSR '18)*, Gothenburg, Sweden, May 27, 2018, pp. 364–375, doi: 10.1145/3196398.3196403.

[24] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza, and R. Oliveto, "Sentiment analysis for software engineering: How far can we go?" in *IEEE/ACM 40th Int. Conf. Software Engineering (ICSE '18)*, Gothenburg, Sweden, May 27, 2018, pp. 94–104, doi: 10.1145/3180155.3180195.

[25] R. Feldman, "Techniques and applications for sentiment analysis," *Comm. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013, doi: 10.1145/2436256.2436274.

[26] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Feb. 2013, doi: 10.1109/MIS.2013.30.

[27] L. E. Duers, "The learner as co-creator: A new peer review and self-assessment feedback form created by student nurses," *Nurse Educ. Today*, vol. 58, pp. 47–52, 2017, doi: 10.1016/j.nedt.2017.08.002.

[28] M. Wen, D. Yang, and C. Rose, "Sentiment analysis in MOOC discussion forums: What does it tell us?" in *Educational Data Mining (EDM '14)*, London, U.K., Jul. 4, 2014.

[29] D. S. Chaplot, E. Rhim, and J. Kim, "Predicting student attrition in MOOCs using sentiment analysis and neural networks," in *Artificial Intelligence in Education Workshop (AIED '15)*, vol. 53, Madrid, Spain, Jun. 2015, pp. 54–57.

[30] G. G. Esparza, A. P. Díaz, J. Canul-Reich, C. A. De-Luna, and J. Ponce, "Proposal of a sentiment analysis model in tweets for improvement of the teaching-learning process in the classroom using a corpus of subjectivity," *Int. J. Combinatorial Optim. Problems and Inform.*, vol. 7, no. 2, pp. 22–34, May 2016.

[31] M. Pousada, S. Caballé, J. Conesa, A. Bertrán, B. Gómez-Zúñiga, E. Hernández, M. Armayones, and J. Moré, "Towards a web-based

[31] teaching tool to measure and represent the emotional climate of virtual classrooms," in *Int. Conf. Emerging Internetworking, Data & Web Technologies*, Wuhan, China, Jun. 10, 2017, pp. 314–327, doi: 10.1007/978-3-319-59463-7_32.

[32] M. Schubert, D. Durruty, and D. A. Joyner, "Measuring learner tone and sentiment at scale via text analysis of forum posts," in *Proc. 8th Edition Int. Workshop Personalization Approaches Learning Environments (PALE)*, London, U.K., Jun. 27–30, 2018.

[33] F. Clarizia, F. Colace, M. De Santo, M. Lombardi, F. Pascale, and A. Pietrosanto, "E-learning and sentiment analysis: A case study," in *Proc. 6th Int. Conf. Information and Education Technology*, London, U.K., Jan. 6, 2018, pp. 111–118, doi: 10.1145/3178158.3178181.

[34] F. F. Balahadia, M. C. G. Fernando, and I. C. Juanatas, "Teacher's performance evaluation tool using opinion mining with sentiment analysis," in *IEEE Region 10 Symposium (TENSYMP '16)*, Sanur, Indonesia, May 9, 2016, pp. 95–98, doi: 10.1109/TENCONSpring.2016.7519384.

[35] S. Rani and P. Kumar, "A sentiment analysis system to improve teaching and learning," *Comput.*, vol. 50, no. 5, pp. 36–43, 2017, doi: 10.1109/MC.2017.133.

[36] J.-a. P. Lalata, B. Gerardo, and R. Medina, "A sentiment analysis model for faculty comment evaluation using ensemble machine learning algorithms," in *Proc. 2019 Int. Conf. Big Data Engineering*, Jun. 11, 2019, pp. 68–73, doi: 10.1145/3341620.3341638.

[37] M. Fleming, B. Coulter, and S. Weaver, "Exploring the student experience of industry placements using sentiment analysis," in *29th Australasian Association for Engineering Education Conf. (AAEE '18)*. Hamilton, New Zealand: Engineers Australia, 2018, p. 213.

[38] T. L. Lynch, "Soft (a) ware in the english classroom: Spreadsheets and sinners: How and why english teachers can claim their rightful place in stem education," *The English J.*, vol. 104, no. 5, pp. 98–101, May 2015.

[39] E. F. Gehringer, "Board 60: Peerlogic: Web services for peer assessment," in *Proc. 2019 ASEE Annu. Conf. & Expo.*, Tampa, FL, USA, Jun. 15, 2019.

[40] C. Gilbert and E. Hutto, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *8th Int. Conf. on Weblogs and Social Media (ICWSM-14)*, Oxford, U.K., May 16, 2014. [Online]. Available: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

[41] Z. J. Beasley and L. A. Piegl, "A domain-dependent lexicon to augment cad peer review," *Comput.-Aided Des. and Appl.*, vol. 18, no. 1, pp. 186–198, May 2021, doi: 10.14733/cadaps.2021.186-198.

[42] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," Jul. 2013. [Online]. Available: arXiv:1307.2579

[43] L. A. Piegl, Z. J. Beasley, and P. Rosen, "Assessing student design work using the intelligence of the crowd," in *Proc. Computer-Aided Design Conf. and Exhibition (CAD '19)*, Changi, Singapore, Jun. 24–26, 2019, pp. 117–121, doi: 10.14733/cadconfP.2019.117-121.

[44] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 49–72, Aug. 2019, doi: 10.1162/tacl_a_00254.

[45] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2019. [Online]. Available: arXiv:1906.08237

[46] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," Mar. 2011. [Online]. Available: arXiv:1103.2903

[47] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. and Trends® Inf. Retrieval*, vol. 2, no. 1–2, pp. 1–135, Mar. 2008, doi: 10.1561/1500000011.

[48] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 35th Annu. Meeting Association Computational Linguistics and 8th Conf. European Chapter Association Computational Linguistics*, Madrid, Spain, Jul. 1997, pp. 174–181, doi: 10.3115/976909.979640.

[49] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Association Computational Linguistics*, Philadelphia, PA, USA, Dec. 11, 2002, pp. 417–424, doi: 10.3115/1118693.1118704.

[50] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM Int. Conf. Knowledge Discovery and Data Mining (SIGKDD '04)*, Seattle, WA, USA, Aug. 22, 2004, pp. 168–177, doi: 10.1145/1014052.1014073.

[51] Z. Beasley, "Crowd-sourced plagiarism detection of essays," in *Proc. Southeast Int. Center for Academic Integrity Conf. (ICAI '20)*, Tampa, FL, USA, Jun. 6–11, 2020. [Online]. Available: https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1002&context=southeast-icai-conference

[52] Z. J. Beasley and L. A. Piegl, "Helps: A domain-specific lexicon for cad peer review," in *Proc. CAD Conf. and Exhibition*, Barcelona, Spain, Jul. 6–8, 2020, pp. 21–25, doi: 10.14733/cadconfP.2020.21-25.

[53] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar, "Building a sentiment summarizer for local service reviews," in *Proc. World Wide Web Workshop NLP Challenges Information Explosion Era (NLPIX '08)*, vol. 14, Beijing, China, Apr. 22, 2008, pp. 339–348.

[54] Z. J. Beasley, "Sentiment analysis in peer review," Ph.D. dissertation, University of South Florida, 2020. [Online]. Available: https://scholarcommons.usf.edu/etd/8160

[55] C. Potts, "On the negativity of negation," *Semantics and Linguistic Theor.*, vol. 20, pp. 636–659, Aug. 2010, doi: 10.3765/salt.v20i0.2565.

[56] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, "A survey on the role of negation in sentiment analysis," in *Proc. Workshop Negation and Speculation in Natural Language Processing*, Uppsala, Sweden, Jul. 2010, pp. 60–68.

[57] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, May 28, 2002, pp. 79–86.

[58] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Sentiful: A lexicon for sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 22–36, Feb. 2011, doi: 10.1109/T-AFFC.2011.1.

[59] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. Human Language Technology Conf. and Conf. Empirical Methods in Natural Language Processing*, Vancouver, Canada, Oct. 2005, doi: 10.3115/1220575.1220619.

[60] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proc. 12th Int. Conf. World Wide Web (WWW '03)*, Budapest, Hungary, May 20, 2003, pp. 519–528, doi: 10.1145/775152.775226.

[61] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J. Big Data*, vol. 2, no. 1, p. 5, Dec. 2015, doi: 10.1186/s40537-015-0015-2.

[62] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Center Research Psychophysiology, Univ. Florida, Gainsville, FL, USA, Tech. Rep. C-1, Jan. 1999.

[63] L. Wu, F. Morstatter, and H. Liu, "Slangsd: Building and using a sentiment dictionary of slang words for short-text sentiment classification," Aug. 2016. [Online]. Available: arXiv:1608.05129

[64] L. Deng and J. Wiebe, "Mpqa 3.0: An entity/event-level sentiment corpus," in *Proc. 2015 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies*, Denver, CO, USA, 2015, pp. 1323–1328, doi: 10.3115/v1/N15-1146.

[65] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," *LREC*, vol. 10, no. 2010, pp. 2200–2204, May 2010.

[66] F. Hoffmann, T. Bertram, R. Mikut, M. Reischl, and O. Nelles, "Benchmarking in classification and regression," *Wiley Interdisciplinary Rev.: Data Mining and Knowl. Discovery*, vol. 9, no. 5, p. e1318, Sep. 2019, doi: 10.1002/widm.1318.

[67] Z. Beasley, A. Friedman, L. Pieg, and P. Rosen, "Leveraging peer feedback to improve visualization education," in *IEEE Pacific Visualization Symposium (PacificVis '20)*, Jun. 3, 2020, pp. 146–155, doi: 10.1109/PacificVis48177.2020.1261.