Capturing Student Feedback and Emotions in Large Computing Courses: A Sentiment Analysis Approach

Marion Neumann
m.neumann@wustl.edu
Washington University in St. Louis
St. Louis, Missouri, USA

Robin Linzmayer
rlinzmay@gmail.com
Washington University in St. Louis
St. Louis, Missouri, USA

ABSTRACT

Enrollment numbers in computer science courses are higher than ever and keep growing. This renders communication and interaction of instructors with individual students extremely challenging, leading to an increase in anonymity (*anonymity gap*). Especially when students struggle in computing courses, personalized help is crucial for them to overcome their problems and frustration and eventually succeed in their studies. At the same time detecting students' misconceptions and gathering feedback at scale is time consuming, resulting in a lack of unbiased feedback available to course instructors (*feedback gap*). Real-time student feedback is a crucial source for instructors to adapt their teaching pace, teaching materials, or course content during the course of the semester to cater to an increasingly diverse student population. In this paper, we investigate a scalable approach to collect and analyze student feedback and emotions. We find that sentiment analysis can efficiently capture student emotions, bearing the potential to lessen both the anonymity and feedback gaps.

CCS CONCEPTS

• **Social and professional topics → Computing education**.

KEYWORDS

student feedback, emotions, sentiment analysis

ACM Reference Format:
Marion Neumann and Robin Linzmayer. 2021. Capturing Student Feedback and Emotions in Large Computing Courses: A Sentiment Analysis Approach. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21), March 13–20, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3408877.3432403

1 INTRODUCTION

Enrollment numbers in computer science (CS) have surged in the United States since 2006 and keep increasing [4]. At the same time demands on graduates, especially in computing, keep growing, leaving students with frustratingly packed course schedules and under enormous academic pressure. Hence, it is no coincidence that mental health issues among computing students are increasing [12, 13, 25, 35]. When class sizes are small, course instructors are likely able to identify students that fall behind or disconnect, and appropriate support can be provided. However, when faculty teach (multiple) high-enrollment courses there is minimal room for personal interaction. This growing anonymity gap between students and instructors makes it impossible to detect when students are frustrated. Real-time feedback on students' learning progress is limited to grade statistics and online feedback forms directly asking about students' experiences typically using Likert-scale questions. Both sources of information are problematic. Likert-scale surveys lack honest or useful feedback as we will discuss in Section 4.1. Grades are often noisy due to group submissions or the fact that grading is performed by multiple graders. Additionally, grades are only available after a delay of often one or two weeks due to time-intense grading processes.

Sentiment analysis (SA) – a reliable and efficient way to recognize emotions in text [20, 34] – is able to overcome these limitations. By applying SA to unit-of-study reflections we are able to gauge all students' emotions automatically. Our hope is to close, or at least limit, both the anonymity and feedback gaps in large courses by providing a mechanism to identify students – as well as course materials – at risk throughout the semester essentially at the time of the assignment submissions. Once identified in a timely manner, instructors can provide personalized help for students at risk, adjust their teaching routines to accommodate students' needs, or adapt course contents. This is especially important in (mostly upper-level) courses covering fast-changing technologies and evolving computing frameworks. Our goal is to create a positive learning experience for diverse student populations in such courses.

As a first step towards this goal, we investigate a simple SA framework that automatically measures, analyzes, and summarizes student feedback and emotions from unit-of-study reflections collected in the form of assignment reviews. This approach has multiple benefits. Data collection may be triggered by a simple prompt on the assignment, instructing students to write a short reflection reporting their experience when working on the assignment. Once the feedback is collected, text analysis is performed automatically to infer an emotion score in real-time. These sentiment predictions can then be analyzed using simple data analysis and graphing tools, such as computing and plotting means and standard deviations or trends across the semester. In addition to easily interpretable summary statistics and aggregated results, free-form textual feedback is available for manual examination or summary visualizations.

We performed an empirical study in a large upper-level computing course to asses the feasibility and usefulness of this approach. In addition to assignment reviews, we collected star ratings on a five-point Liket scale, assignment grades, and demographic information to analyze emotions across different student populations.

This work is licensed under a Creative Commons Attribution International 4.0 License.

SIGCSE '21, March 13–20, 2021, Virtual Event, USA.
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8062-1/21/03.
https://doi.org/10.1145/3408877.3432376

We further gathered sentiment labels for each written review from three human annotators to validate the sentiment predictions.

1.1 Research Questions

To understand whether capturing student feedback and emotions is feasible, as well as whether their analysis provides useful insights for course instructors, our study was driven by the following two main research questions:

RQ1: Can we use sentiment analysis (SA) to efficiently measure student emotions from free-form textual feedback?

RQ2: Do these predicted student emotions provide *valuable* insights for course instructors?

Since RQ2 is very broad, we specifically investigate the following three more focused questions:

2-P: *Performance*: Is there, if at all, a correlation of students' emotions and their assignment grades?

2-D: *Demographics*: Is there a difference in emotions for different student populations?

2-T: *Trend/Content*: Do emotions change over time or based on the assignment topic?

Hence, *valuable* insights for instructors as referred to in RQ2 could for instance be that some student populations have a more negative experience when working on their assignments than others or that certain topics or semester times trigger more negative emotions.

2 RELATED WORK

We will review previous work investigating student emotions especially in CS courses, how they relate to learning, self-efficacy, and performance, and how they can be measured.

2.1 Studying Student Emotions

Student emotions and motivation, and how they relate to academic achievement, is widely studied in both educational psychology [23, 28, 30] and neuroscience [14]. All of these studies report meaningful relationships between emotions and student achievements, indicating that maintaining healthy or positive emotions of students is beneficial for their learning outcomes. Surprisingly, there is not much work in the CS education community on this topic.

Closest related to our work is the quantitative study performed by Lishinski et al. They study the connection between students' emotional reaction to programming projects in CS1 using a basic emotional reactions survey and assignment scores [19]. Other work studies the relationship of emotional health and student success based on a questionnaire and exam scores [7], where no correlation was found. On a much smaller scale a study by Kinnunen and Simon revealed that the primary reflective experience with programming assignments was emotional. A series of quantitative studies using interviews established that students made self-efficacy assessments as a result of their emotions [16–18].

2.2 Measuring Emotions At Scale

Most measuring instruments to assess student emotions take the form of surveys or interviews, both of which need to be carefully crafted based on its intended purpose [11, 18, 19, 27]. Of these, only surveys scale to large-enrollment courses. Another scalable way to assess emotions is via student observation [5], which is often not feasible due to technical challenges and privacy considerations.

Sentiment analysis on free-form feedback has been used as a scalable and easy-to-use alternative. Most previous work focuses on comparing different supervised sentiment analysis algorithms [2, 3, 10]. Being based on labeled training data and fine-tuned text preprocessing and parameter settings the results are usually not easily generalizible to new data. In [31] a simple rule-based approach is explored on a dataset of student comments in a Coursera course as well as on a dataset of comments and ratings for lecture and lab sessions after midterm and final exams. Other previous work analyzes Twitter and social media data [29], longitudinal trends in MOOCs [8], or teaching performance evaluation data [1, 6]. A few studies examine real-time feedback during the course of the semester in form of unit-of-study evaluations [22] or learning diaries [24]. The former focuses on comparing different SA approaches. The goal of the latter is similar to our work since they aim at improving the communication between students and instructors to enhance students' learning experience. However, the authors only report a proof of concept analysis showing that their system successfully presents information in an easy to understand manner and that emotions expressed in learning diaries can be extracted in a meaningful way using a small dataset of 105 diary entries [24].

Table 1: Demographics of the 63 participants.

Gender	Ethnicity	Standing	Major
52% Male	66% Asian/Asian Am.	35% UG	58% CS
26% Female	12% Non-Hispanic White	43% Masters	20% other
22% no resp.	22% no resp.	22% no resp.	22% no resp.

3 METHODS

Similar to the work of Kinnunen and Simon, our study analyzes students' perceptions of their experiences when working on an assignment and not their actual experiences. Perceptions and perceived emotions are important since they influence the students' views on their own abilities and self-efficacy [18].

3.1 Study Setting and Data

The data for our study was collected from students enrolled in an upper-level undergraduate course on cloud computing in Fall 2018 at a North American research-focused institution with institutional approval to study human subjects. Study participation was voluntary, and from the 99 enrolled students 63 participated in the study. Even though the number of students participating in our study is not overly large, the dataset we study is larger since we look at the students' reflections for each of the course's nine assignments.

Further, the participants come from a diverse student population. The students are a mix of junior and senior CS undergraduates (31%), as well as students in CS, engineering, and business masters programs (43%). Gender balance is roughly 2:1 male-to-female. Table 1 summarizes the **demographics** of the study participants based on self-reported data from a survey. 22% of the students did not provide any demographic information; their data is not used in the demographics analysis (2-D). It is included in all other analyses.

The main source of data for our study are 518 **homework reviews**. These reviews are self-reported unit-of-study reflections that were collected at the same due date as each of the nine homework assignments in exchange for bonus points. Since one of the course topics introduces sentiment analysis as an application of big data analytics, we asked the students to provide a "product review" for each completed assignment. In addition to the instructions to write at least 50 words, we used the following prompt:

> *We give you the chance to voice your opinions on each homework by writing a couple of sentences as a review for the homework. You will not be graded on what your review says, but solely for the completion of it. At the end of the year, given that we have enough data, you will perform sentiment analysis on these reviews to see which assignments you and your peers regarded as "positive" and which as "negative".*

Next to the textual reviews, we also collected student-reported **star ratings** for each assignment using a five-star scale. Each student was asked to provide a rating for their own assignment, again viewing the homework as a product and rating it as part of their review process similar to what customers would do on an e-commerce platform such as `amazon.com`. Note that there might be a discrepancy between the self-reported star ratings and the emotions voiced in the review. We will discuss this in more detail in Section 4.

To account for this issue and also that self-reported ratings might be subjective or biased, we asked three human annotators – two of which were students that were not part of the study team of which one had taken the course in a previous semester – to provide a star rating after reading each of the 518 textual statements. We will use the median of those three ratings as **external rating** for each review. Both student-reported and the median externally-annotated ratings will be used as ground truth to study the sentiment prediction approach when investigating RQ1.

Further, we will be using **assignment grades** on a scale from 0-100% to asses performance achievements.

3.2 Sentiment Analysis Approach

The main goal of sentiment analysis (SA) is to automatically *infer* emotions from free-form text documents, such as product reviews or tweets, and then *use* these emotions, for example, to build a recommendation system or to gauge political opinions [34].

3.2.1 Overview. There are two basic inference tasks: polarity detection and emotion recognition [9]. The former distinguishes two binary classes, where each piece of text typically stating an opinion on a single issue is classified as one of two opposing sentiments such as "positive" vs. "negative". The latter is often modeled as *rating prediction* using a discrete scale, such as number of stars, or a continuous score in $[-1, 1]$. The two major ways to approach both tasks are supervised and unsupervised SA methods. Supervised SA uses the classical supervised machine learning approach [26], where a classification or regression model is trained on labeled training examples, each being a text document associated with a ground truth polarity class label or star rating. The most common unsupervised approach is rule-based SA, where a lexicon of sentiment expressions is used to determine a document's sentiment score. Unsupervised approaches have two main advantages, they are less domain sensitive and they do not require costly to obtain training labels. See [20] for an in-depth introduction of SA approaches.

3.2.2 VADER. For our study, we perform rating prediction using VADER (Valence Aware Dictionary and sEntiment Reasoner), a recently introduced, simple rule-based method [15], whose main component is a gold-standard sentiment lexicon especially attuned to microblog-like contexts. We deemed it especially suitable for our purpose, since it is unsupervised and our homework reviews are short student-generated reflections voicing sentiments that are similar to those expressed in social media. Each of the 7517 sentiment terms in the lexicon, such as "admirable", "desperate", or "hard", is associated with 10 independently assigned valence scores. VADER produces a **sentiment score** for each input text document, here our homework reviews, $s_i^{\text{VADER}} = f(r_i)$, where r_i is the preprocessed ith review and s_i^{VADER} is its weighted composite score computed by summing the average valence scores of each lexicon word in r_i, adjusted according to some heuristic rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). Text preprocessing is extremely lightweight, consisting of simple tokenization, lowercasing, and stopword removal. We refrained from performing more elaborate processing such as stemming or negation handling for simplicity. VADER's heuristic rules go beyond what is captured in a typical rule-based sentiment model. It incorporates, for instance, word-order sensitive relationships between terms as well as intensifiers, booster words, or degree adverbs which impact the sentiment intensity [15]. Moreover, VADER has two advantages that are important for our work. It is computationally economical and it does not need labeled training data in form of annotated reviews as required by supervised approaches. To compute s_i^{VADER} we used the `compound_score` produced by version 3.2.1 of the open-source VADER implementation in Python.[1]

3.3 Data Analysis

Once we have the means to automatically quantify emotions from textual feedback, we can validate, analyze, summarize, and visualize them. In this study, we performed data analysis on unit-of-study reviews and numeric emotion scores using quantitative statistical methods, graph visualizations, and tools to visualize free-form text.

To visualize the content of the reviews we used the Python wordcloud package.[2] Each wordcloud visualizes the top 15 terms in a given set of reviews according to their frequency and a color coding is used to indicate the sentiment given by the VADER lexicon. To compare the predicted VADER sentiment scores with the student-reported (and externally-annotated) ratings, we rescaled and discretized the VADER scores to $[1, 2, 3, 4, 5]$ and report mean absolute differences (MAD). To assess whether there is a difference in emotions and assignment grades (2-P), we report the correlation coefficient and plot the assignment grade versus sentiment score. To investigate the differences in emotions of various student populations (2-D), we look at VADER scores grouped by demographic features such as gender, major, ethnicity, and academic standing. We performed a two-sided Mann-Whitney U test and report statistically significant differences in sentiment scores on an $\alpha = 0.05$ significance level. Further, we compare the distributions of assignment grades g_i, VADER scores s_i^{VADER}, and their difference $(g_i - s_i^{\text{VADER}})$

[1] https://github.com/cjhutto/vaderSentiment
[2] https://amueller.github.io/word_cloud

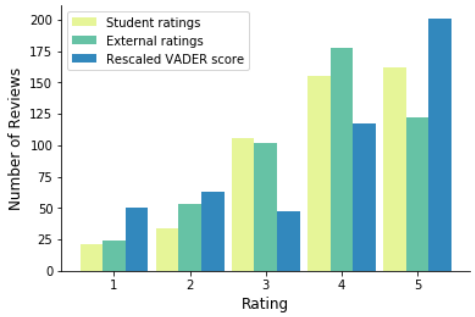

Figure 1: Distributions of student self-reported star ratings, externally annotated ratings, and rescaled VADER scores.

grouped by gender. To assess whether these distributions are significantly different, we used the Anderson-Darling test [33]. To analyze trends across the semester (2-T) we plot average VADER scores and standard errors for each assignment grouped by demographics. We used Python 3.7.3 for all analyses.

4 VALIDATING SENTIMENT PREDICTIONS

In this section, we will discuss common drawbacks of star ratings and investigate whether SA using the VADER model is able to measure student emotions in a meaningful and efficient way (RQ1).

4.1 Challenges with Star Ratings

Star ratings are an implementation of the Likert scale which is typically used to scale responses in survey research. The most frequently used Likert-type scale is five [21], which is also most commonly used for product or movie reviews. The main obvious drawback of star ratings is that no explanation on why something is liked or disliked is provided. Further, star ratings are often extremely noisy, their general reliability is questionable, and positive ratings are typically overrepresented [32]. Figure 1 shows this behavior for our data; students reported less 1- and 2-star ratings compared to both the external annotators and the VADER scores. Possible explanations could be cultural differences (66% of students self-reported as Asian/Asian American) or that students worry that their grade will be affected by giving a negative rating. When using an odd number of choices the middle option provides an easy way for the rater to not take sides. In our study, both student and external annotators are more inclined to give the neutral rating instead of making a decision in favor of one of the polar sides. Figure 1 shows that rescaled VADER scores in the 3-star bin occur only half as often than self-reported or externally-annotated ones. It turns out that 3-star reviews can be anything from extremely negative to extremely positive. One explanation of this behavior is that 3 stars are used to mean "undecided" rather than a neutral emotion. This is undesirable when using emotions to identify students or course materials at risk. Hence, we conclude that ratings on a five-star scale are not suitable to assess student emotions.

4.2 Sentiment Terms

In the following, we investigate whether soliciting textual feedback in the form of reviews and predicting sentiment scores is a feasible alternative. As a sanity check we visualize the 15 most frequent

Figure 2: Most frequent sentiment terms in all positive vs. negative reviews. Larger means more frequent. Green (pink) terms have a positive (negative) value in the VADER lexicon.

sentiment terms across all reviews with a negative versus a positive VADER score. Figure 2 shows that positive reviews contain mainly positive terms with a couple of less frequent negative terms such as "difficult" and "confusing". In reviews with a negative VADER score the most frequent terms are "hard" and "difficult". We also see other negative terms such as "frustrating" and "unclear", which is what we would expect in course assignment reviews. Note that more elaborate text preprocessing such as stemming could be considered to consolidate frequencies for words such as "helps" and "help".

4.3 VADER vs. Star Ratings

Next, we compare VADER scores and student-reported ratings. In our data, we observe a lot of reviews with low VADER score but high student rating in addition to 3-star ratings extending across the entire VADER range. Only for reviews labeled with 1- and 2-stars we see the expected behavior: most negatively rated reviews have a low VADER score. To get a better sense of whether VADER or ratings provide a better estimate of the sentiment voiced in the reviews, we looked at some examples (omitted due to space limitations).

For the reviews with low VADER score but high student rating the text typically clearly reveals that the sentiment is negative, agreeing with the VADER score. So, why does VADER fail for some of the low rated reviews? Looking at the text shows that these reviews contain expressions with positive connotation in a negative or neutral context. For instance "spark" is considered a positive term according to VADER, but in the context of this course it refers to a cloud computing framework, which shouldn't have an emotion score. This shows that VADER predictions are not perfect; however, while scanning the actual reviews, we only found few such cases.

How do VADER scores relate to ratings provided by independent human experts? Table 2 reports the pairwise mean absolute differences between VADER scores, externally-annotated ratings, student-reported ratings, and randomly generated ratings (RAND). Whereas the student ratings are closest to the external ratings

Table 2: Mean absolute difference (MAD) of sentiment scores (VADER), externally-annotated (external), student-reported (student), and randomly-generated (RAND) ratings.

	Mean absolute difference (MAD)			
	VADER	external	student	RAND (*)
VADER	–	0.95	1.14	1.70
external	–	–	0.67	1.53
student	–	–	–	1.58

(*) averaged over 20 independent random samples

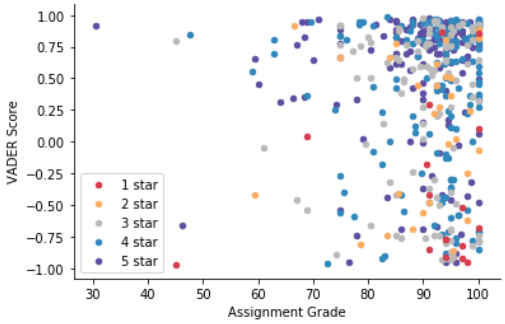

Figure 3: Assignment grades from 30% to 100% vs. VADER scores from -1 to 1 colored by star rating (1-5 stars).

with an MAD of 0.67, the first row shows that the VADER scores are on average closer to the external ratings (MAD of 0.95) than to the student-reported ones (MAD of 1.14). The last column, in turn, shows that both student and external ratings are closer to randomly generated ratings (MAD of 1.58 and 1.53) than VADER scores.

These findings show that if we cannot afford human-annotated sentiment scores, automatically predicted VADER scores offer a good alternative to time-consuming manual reading and labeling.

5 RESULTS

Now that we have established that SA can be used to efficiently measure student emotions, we will instigate whether these emotions provide valuable insights for course instructors (RQ2).

5.1 Performance (2-P)

Plotting the assignment grades versus the VADER sentiment score grouped by student-reported star rating, cf. Figure 3, shows that there is no obvious correlation between grades and emotions. In fact, the correlation coefficient is 0.0. This shows that sentiment scores provide orthogonal information to assignment grades, which are often the only measure used to assess students' learning experience during the course of a semester. Hence, we conclude that considering student emotions is extremely important for course instructors to get a holistic view of their students' course experience.

5.2 Demographics (2-D)

Table 3 shows the differences in VADER scores grouped by gender, academic standing, ethnicity, and major. We see a statistically significant difference in emotions for male versus female students under the two-sided Mann-Whitney U test with a significance level of $\alpha = 0.05$. Female students have more negative reviews, indicating that this student population reports more critical towards their homework experience. They use terms like "difficult" and "hard" more frequently than their male peers and they also mention that things are "unclear" and "confusing". When considering academic standing, graduate students have slightly more negative emotions than undergraduates. Comparing CS majors versus others we also see a slight difference; CS majors are more critical/negative toned than business and engineering students. However, neither difference is statistically significant.

Triggered by these gender differences in emotions and previous studies on gender differences in CS in general [19], we performed a more detailed comparison of emotions of female vs. male students in our study data. When comparing the emotions students had when working on the assignments with the grades they achieved, we observe that a larger fraction of female students have significantly higher grades and a larger fraction of male students have significantly higher emotions as shown in Figure 4 (*middle* and *right*). This means that more female students have high assignment grades but lower emotion scores as shown in the left panel of Figure 4. The density of the difference of (normalized) grades and emotion scores is higher for females in the range $[0, 1]$. All three distributions (grades, sentiment scores, and their difference) are significantly different under the Anderson-Darling test ($\alpha = 0.05$) with p-values of 0.0041, 0.0036, and 0.0005 respectively. Due to space constraints we have to leave a more detailed discussion of those results for future work. Overall, we can conclude that emotions differ for different student populations especially when considering gender, which answers 2-D affirmatively.

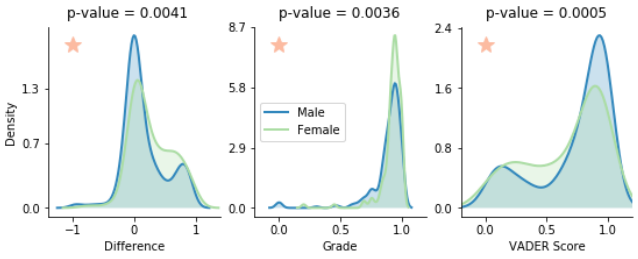

Figure 4: Normalized assignment grades, VADER scores, and their difference by gender. Stars indicate significantly different distributions under the Anderson-Darling test ($\alpha = 0.05$).

5.3 Trends (2-T)

Figure 5 visualizes VADER scores grouped by gender, academic standing, and major for each assignment. Looking at the average trend we can see a slight drop in emotion scores for hw5 and hw9, whereas the highest scores occur for hw6 to hw8. Both observations make sense given the covered course content and times in the semester. Students are likely more busy in the middle and at the end of the semester for midterms and final projects. These times correspond

Table 3: Differences in emotion scores by demographics. We report average VADER scores ($\bar{s}^{\text{VADER}}$) per group, their standard errors ($\varepsilon^{\text{VADER}}$), and the p-values of a two-sided Mann-Whitney U test; significant differences ($\alpha = 0.05$) in bold.

	groups	# of reviews	$\bar{s}^{\text{VADER}}$	$\varepsilon^{\text{VADER}}$	p-value
Gender	Male	236	0.44	0.04	**0.0034**
	Female	135	0.31	0.05	
Ethnicity	Asian	311	0.39	0.03	0.4707
	White	60	0.40	0.08	
Standing	UG	179	0.44	0.04	0.1662
	Masters	192	0.35	0.05	
Major	CS	276	0.39	0.04	0.2889
	other	87	0.42	0.07	

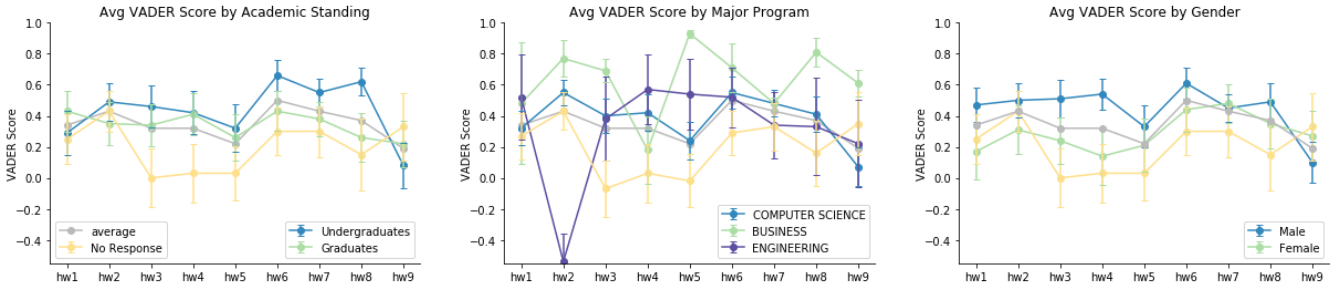

Figure 5: Emotion score computed for all homework reviews grouped by assignment and student population (standing, major, and gender). Some students wrote homework reviews, but did not provide demographic information (labeled *No Response*).

to the feedback given for hw5 and hw9 respectively. The course for which we collected the data covers cloud computing. The first half of the semester introduces the concepts and framework in a more theoretical way, whereas the second half of the semester becomes more applied covering specific Hadoop tools and big data applications. hw6 and hw8 being rated highest shows that students tend to like those materials better. Interestingly, this trend is more pronounced for undergraduate students who seem to enjoy working on the applied topics (hw6-hw8) more than their graduate peers.

Comparing CS majors versus others we see that business students tend to have more positive emotions than computer science and engineering students. One possible explanations could be that there is a self-selection bias in that business students enrolling in computer science classes are more motivated and interested in hands-on applied computing knowledge. However, this is just an educated guess which should be investigated further. Analyzing trends by gender reveals that emotion scores of female students are lower than male students towards the beginning of the semester and for assignments with major programming focus (hw4, hw8).

Another interesting observation is that students that did not provide demographics (coded as *No Response*) have lower sentiment scores than those who self-reported demographics. Last, these visualizations may reveal useful information about the difficulty of the assignments. Overall, we find that emotions differ for different course topics and semester times, which answers 2-T affirmatively.

6 THREATS TO VALIDITY

One threat to the validity of this study is that the reviews were collected in exchange for bonus points (5% of the assignment grade). However, given that the dataset contains a total of 518 reviews and almost all students submitted the feedback throughout the semester, we expect minimal effects of bias through self-selection.

A common threat to validity in educational research that require institutional review board approval are small sample sizes, which may hinder meaningful statistical analysis. Our dataset comprises data from 63 students from one course in one institution. However, most of our quantitative analysis are based on a dataset of 518 text documents, one review per student and homework assignment. When considering some of the challenges to set up such studies in upper-level college courses, such as getting students to participate in a meaningful and honest manner, our study population size is still reasonable. In the future, we aim to study RQ2 on a larger scale to mitigate potential population bias by extending data collection to other courses and institutions. We currently collect reflections in our introduction to computer science course with an enrollment of over 300 students.

A technical threat is posed by the fact that VADER is specifically attuned to sentiments in social media. This might cause the sentiment scores to be skewed, and if the approach were adapted to the educational setting the scores might be slightly different. However, we consciously chose to use the out-of-the-box method to facilitate emulation. The idea is that course instructors can easily get analysis results based on reflections and extracted emotions that are complementary to performance scores. When used in an early warning system to identify struggling students, we recommend fine-tuning the VADER lexicon to the specific course setting.

7 CONCLUSIONS AND FUTURE WORK

We studied sentiment analysis as an effective approach to identify student emotions from free-form textual feedback. Our study shows that SA can extract meaningful emotions (RQ1), and that these emotions provide valuable insights for course instructors that go beyond the students' grades or self-reported star ratings (RQ2).

Specifically, we show that VADER, a simple rule-based sentiment predictor, produces emotion scores that are close to those of human annotators. Since the goal of this paper was to investigate whether it is *feasible* and *useful* to analyze student emotions in large computing courses, we leave the tasks of fine-tuning the VADER lexicon to the specific course setting, as well as comparing different sentiment predictors and finding the best one as future work.

Even though only 63 students participated in our study, we were able to base most results on the analysis of 300-500 text documents. We are planning to extend our study to student populations in other courses including CS1 and other institutions in the future.

When analyzing the predicted student emotions, we found interesting differences for various student populations, across semester times, and for different assignment topics. Providing these insights to course instructors during the course of the semester may help to decrease the *feedback gap* between students and instructors. Gender differences in emotions and the lack of correlation between emotions and assignment grades are particularly interesting findings of our study that should be investigated further in future work.

We also think that sentiment analysis is a promising tool to incorporate emotions into early warning systems to identify struggling students to limit the *anonymity gap* between students and course instructors in large computing courses.

ACKNOWLEDGMENTS

This work is under the oversight of the IRB at Washington University in St. Louis. We would like to thank Zach Mekus and Kevin Fu for reading and labeling the reviews. This material is based upon work supported by the NSF under Grant Nos. 1525028, 1525173, 1525373 (DEERS project).

REFERENCES

[1] Paola Adinolfi, Ernesto D'Avanzo, Miltiadis D Lytras, Isabel Novo-Corti, and Jose Picatoste. 2016. Sentiment analysis to evaluate teaching performance. *International Journal of Knowledge Society Research (IJKSR)* 7, 4 (2016), 86–107.

[2] Nabeela Altrabsheh, Mihaela Cocea, and Sanaz Fallahkhair. 2014. Learning sentiment from students' feedback for real-time interventions in classrooms. In *International Conference on Adaptive and Intelligent Systems*. Springer, 40–49.

[3] Nabeela Altrabsheh, Mihaela Cocea, and Sanaz Fallahkhair. 2014. Sentiment analysis: towards a tool for analysing real-time students feedback. In *2014 IEEE 26th international conference on tools with artificial intelligence*. IEEE, 419–423.

[4] Computing Research Association et al. 2017. Generation CS: Computer Science Undergraduate Enrollments Surge Since 2006.

[5] Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

[6] Francis F Balahadia, Ma Corazon G Fernando, and Irish C Juanatas. 2016. Teacher's performance evaluation tool using opinion mining with sentiment analysis. In *2016 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 95–98.

[7] Jens Bennedsen and Michael E Caspersen. 2008. Optimists have more fun, but do they learn better? On the influence of emotional and social factors on learning introductory computer science. *Computer Science Education* 18, 1 (2008), 1–16.

[8] Ida Camacho and Ashok Goel. 2018. Longitudinal trends in sentiment polarity and readability of an Online Masters of Computer Science course. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. ACM, 21.

[9] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems* 28, 2 (2013), 15–21.

[10] V Dhanalakshmi, Dhivya Bino, and AM Saravanan. 2016. Opinion mining from student feedback data using supervised learning algorithms. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*. IEEE, 1–5.

[11] Brian Dorn and Allison Elliott Tew. 2015. Empirical validation and application of the computing attitudes survey. *Computer Science Education* 25, 1 (2015), 1–36.

[12] Bruce Gerson. 2019. Life@CMU Measures Student Stress, Depression. https://www.cmu.edu/news/stories/archives/2019/april/life-at-cmu-project.html [Online; viewed 24-June-2019].

[13] Julie Gould. 2014. Mental health: Stressed students reach out for help. *Nature* 512, 7513 (2014), 223–224.

[14] Christina Hinton, Koji Miyamoto, and BRUNO Della-Chiesa. 2008. Brain research, learning and emotions: implications for education research, policy and practice 1. *European Journal of education* 43, 1 (2008), 87–103.

[15] Clayton J Hutto and Eric Gilbert. 2014. Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth international AAAI conference on weblogs and social media*.

[16] Paivi Kinnunen and Beth Simon. 2010. Experiencing programming assignments in CS1: the emotional toll. In *Proceedings of the Sixth international workshop on Computing education research*. ACM, 77–86.

[17] Päivi Kinnunen and Beth Simon. 2011. CS majors' self-efficacy perceptions in CS1: results in light of social cognitive theory. In *Proceedings of the seventh international workshop on Computing education research*. ACM, 19–26.

[18] Päivi Kinnunen and Beth Simon. 2012. My program is ok–am I? Computing freshmen's experiences of doing programming assignments. *Computer Science Education* 22, 1 (2012), 1–28.

[19] Alex Lishinski, Aman Yadav, and Richard Enbody. 2017. Students' emotional reactions to programming projects in introduction to programming: Measurement approach and influence on learning outcomes. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*. ACM, 30–38.

[20] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[21] Luis M Lozano, Eduardo García-Cueto, and José Muñiz. 2008. Effect of the number of response categories on the reliability and validity of rating scales. *Methodology* 4, 2 (2008), 73–79.

[22] Sunghwan Mac Kim and Rafael A Calvo. 2010. Sentiment Analysis in Student Experiences of Learning.. In *EDM*. 111–120.

[23] Carolina Mega, Lucia Ronconi, and Rossana De Beni. 2014. What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of educational psychology* 106, 1 (2014), 121.

[24] Myriam Munezero, Calkin Suero Montero, Maxim Mozgovoy, and Erkki Sutinen. 2013. Exploiting sentiment analysis to track emotions in students' learning diaries. In *Proceedings of the 13th Koli Calling International Conference on Computing Education Research*. ACM, 145–152.

[25] Julia Nguyen. 2014. The Pressures of Success in Undergraduate Computer Science Programs. https://modelviewculture.com/pieces/the-pressures-of-success-in-undergraduate-computer-science-programs [Online; viewed 24-June-2019].

[26] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86.

[27] Dale Parsons, Krissi Wood, Joy Gasson, and Adon Moskal. 2019. Development of a Self-Reporting Tool for Capturing Student Emotions During Programming Activities. In *Proceedings of the Twenty-First Australasian Computing Education Conference*. ACM, 64–68.

[28] Reinhard Pekrun, Andrew J Elliot, and Markus A Maier. 2009. Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of educational Psychology* 101, 1 (2009), 115.

[29] Fahmi Candra Permana, Yusep Rosmansyah, and Atje Setiawan Abdullah. 2017. Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media. In *Journal of Physics: Conference Series*, Vol. 893. IOP Publishing, 012051.

[30] Elizabeth R Peterson, Gavin TL Brown, and Miriam C Jun. 2015. Achievement emotions in higher education: A diary study exploring emotions across an assessment event. *Contemporary Educational Psychology* 42 (2015), 82–96.

[31] Sujata Rani and Parteek Kumar. 2017. A sentiment analysis system to improve teaching and learning. *Computer* 50, 5 (2017), 36–43.

[32] Diana Schindler, Lars Lüpke, and Reinhold Decker. 2014. Estimating true ratings from online consumer reviews. In *German-Japanese Interchange of Data Analysis Results*. Springer, 235–252.

[33] Michael A Stephens. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association* 69, 347 (1974), 730–737.

[34] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*.

[35] Rachel Zimmerman. 2012. MIT 'Meltdown' Blog Resonates With Stressed-Out Students. https://www.wbur.org/commonhealth/2012/11/09/mit-meltdown-blog-stressed-students [Online; viewed 24-June-2019].