SETSum: Summarization and Visualization of Student Evaluations of Teaching

♠Yinuo Hu* ♠Shiyue Zhang* ♣♡Viji Sathy ♣♡A. T. Panter ♠Mohit Bansal
♠Department of Computer Science, UNC Chapel Hill
♣Department of Psychology and Neuroscience, UNC Chapel Hill
♡Office of Undergraduate Education, UNC Chapel Hill
{huyinuo, shiyue, mbansal}@cs.unc.edu;
{viji.sathy, panter}@unc.edu

Abstract

Student Evaluations of Teaching (SETs) are widely used in colleges and universities. Typically SET results are summarized for instructors in a static PDF report. The report often includes summary statistics for quantitative ratings and an unsorted list of open-ended student comments. The lack of organization and summarization of the raw comments hinders those interpreting the reports from fully utilizing informative feedback, making accurate inferences, and designing appropriate instructional improvements. In this work, we introduce a novel system, SETSum, that leverages sentiment analysis, aspect extraction, summarization, and visualization techniques to provide organized illustrations of SET findings to instructors and other reviewers. Ten university professors from diverse departments serve as evaluators of the system and all agree that SETSum help them interpret SET results more efficiently; and 6 out of 10 instructors prefer our system over the standard static PDF report (while the remaining 4 would like to have both). This demonstrates that our work holds the potential of reforming the SET reporting conventions in the future.

1 Introduction

Colleges and universities rely on student evaluations of teaching (SETs) to assess students' perceptions about their courses (Chen and Hoshower, 2003; Zabaleta, 2007). These evaluations about the course consist of both quantitative ratings using Likert-type scales and open-ended comments that describe student experiences. In many universities, SETs are a standard component of evaluations of teaching and have multiple functions. First, they help individual faculty members examine their own teaching performance in a diagnostic way so they can work to improve their approach in subsequent offerings of the course. Second, SETs allow institution leaders to review and describe the educational quality of course offerings and the performance of instructors. Third, though controversial, SET summaries often are used as part of an instructor's larger portfolio to demonstrate their teaching history during high-stakes settings. Finally, in some colleges and universities, SET summaries are released to students to help guide them with course selections. Given this wide range of uses for the SET summaries, it is important that thoughtful, accurate, and well-designed representations are provided to draw accurate inferences about teaching quality, course design, and student learning (see ethics Sec. 7 for more details).

Usually, at the end of each semester, SET results are summarized into a PDF report for instructors or other reviewers. As shown in a sample standard SET report in Fig. 5 of Appendix, quantitative ratings are summarized using basic statistics, such as mean and median, while students' comments from open-ended questions are simply listed as raw text – without adequate organization and analyses. When a college course is particularly large (e.g., with more than 100 students), the final SET report can be longer than 10 pages, which is time-consuming to read and analyze (Alhija and Fresko, 2009). In addition, instructors' or other reviewers' own cognitive biases may lead to inaccurate inferences and analyses, e.g., people tend to pay more attention to negative than positive comments (Kanouse and Hanson Jr, 1987).

Therefore, the goal of our work is to provide a new dynamic presentation of SET results to facilitate more efficient and less biased interpretations compared to the standard PDF report. After obtaining institutional SET data of four semesters from the University of North Carolina (UNC) at Chapel Hill, for demonstration we select two quantitative and two open-ended questions from the total number of questions (Sec. 3). We develop a system, SETSum, to summarize and visualize the results of these four questions. For quantitative ratings

[*] Equal contribution.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, pages 71 - 89
July 10-15, 2022 ©2022 Association for Computational Linguistics

(Sec. 4.1), we visualize two statistics: *response rate* and *sentiment distribution*. For open-ended comments (Sec. 4.2), we develop a sentiment analysis model to predict whether each comment sentence is positive or negative. We use an aspect extraction approach to help instructors quickly know the popularly discussed topics by students, e.g., assignment, and the corresponding topic sentiments. Finally, we propose an unsupervised extractive summarization method that extracts top sentences with high centrality, low redundancy, and balanced sentiments as a summary of each aspect.

Automatic evaluations (Sec. 5.1) demonstrate that our sentiment prediction and aspect extraction modules achieve good accuracy, and our summarization method produces more diverse and less biased summaries than simply picking top central sentences. More critically, the effectiveness of SETSUM should be judged by its main users – instructors. Thus, we begin by conducting human evaluations (Sec. 5.2) with 10 professors from 8 different academic departments at UNC. Note that SETSUM is continuously under development, and our human evaluations were conducted on our very first version: SETSUM V1.0. After evaluating the two SET presentation approaches, instructors are asked to complete a survey comparing the usefulness of SETSUM to the standard SET report. According to their responses, most of the new features introduced on SETSUM are perceived as *useful* to *very useful* by most instructors (on average, 8.8 out of 10), compared to the standard report. All 10 instructors agree that SETSUM helps them interpret their ratings and comments more efficiently; while 4 out of 10 think the new system also supports less biased interpretations. Finally, 6 of 10 favor SETSUM more than the standard approach; the remaining 4 think both reports could be helpful. Overall, for our first evaluation, instructors hold a *positive* attitude towards SETSUM and offer valuable and constructive suggestions to us.

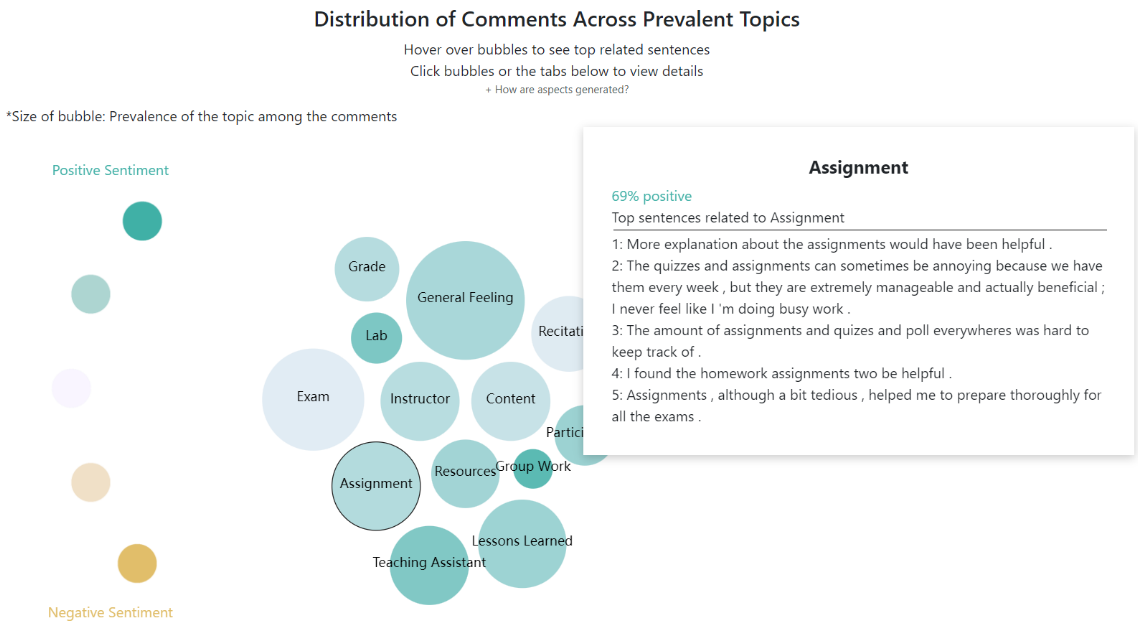
Lastly, in Sec. 7, we discuss if machine-involved representations of SETs may introduce new errors or bias and if so, what improvement needs to be made before the "demonstration" can transition to an "application". Our system aims to provide accurate, efficient, and visualized SET results to instructors or other reviewers. It does not directly make any value judgments or evaluations about the instructor's skills, the course design, or the amount of student learning during the term.

To the best of our knowledge, despite its widespread use, we are among the few researchers to develop a pilot system that presents student-reported evaluations of teaching by using natural language processing (NLP) techniques. In addition, we are the first to apply the system for a SET instrument and evaluate it using actual SET data from a large public university. Though more development work is in progress, our results demonstrate that our approach is promising to reform the SET report conventions in the future. Our SETSUM V1.1 website requires credentials to login, please contact us for an access to the website. We provide a YouTube video to walk you through SETSUM V1.1. Our code is hosted at SETSum Github Repo.

2 Background & Related Work

As mentioned, SETs are widely used in higher education (Chen and Hoshower, 2003; Zabaleta, 2007). SET studies have shown that they can capture students' opinions about instruction (Balam and Shannon, 2010), enhance course design, can be used as a tool for assessing teaching performance (Penny and Coe, 2004; Chen and Hoshower, 2003), and reflect institutional accountability about teaching (Spooren et al., 2013). Many instructors view SETs as valuable feedback to improve their teaching quality (Griffin, 2001; Kulik, 2001). Many studies focus on instrument design (i.e., which questions to ask), reliability and validity of SET results (i.e., are the scores consistent across contexts; are scores related to other key constructs), and potential confounding variables that affect SETs (e.g., do scores differ by discipline, instructor race/ethnicity and gender, student grade) (Simpson and Siguaw, 2000; Spooren et al., 2013).

Typical SET instruments include quantitative Likert-scale ratings. They are supplemented by open-ended comments (Stupans et al., 2016; Marshall, 2021). Therefore, compared to quantitative ratings, open-ended comments are often under-analyzed or ignored completely due to labor required to provide an adequate summary (Alhija and Fresko, 2009; Hujala et al., 2020), raising the need for contemporary methods in automated text analysis. Recent works start to analyze student comments via text mining and machine learning methods such as sentiment analysis (Wen et al., 2014; Azab et al., 2016; Cunningham-Nelson et al., 2018; Baddam et al., 2019; Sengkey et al., 2019; Hew et al., 2020), and identify *topics*, *themes*, or



Distribution of Comments Across Prevalent Topics

Hover over bubbles to see top related sentences

Click bubbles or the tabs below to view details

+ How are aspects generated?

*Size of bubble: Prevalence of the topic among the comments

Assignment

69% positive

Top sentences related to Assignment

1: More explanation about the assignments would have been helpful .

2: The quizzes and assignments can sometimes be annoying because we have them every week , but they are extremely manageable and actually beneficial ; I never feel like I 'm doing busy work .

3: The amount of assignments and quizes and poll everywheres was hard to keep track of .

4: I found the homework assignments two be helpful .

5: Assignments , although a bit tedious , helped me to prepare thoroughly for all the exams .

Figure 1: The distribution of comments across prevalent topics (aspects). On the left, it shows the aspect bubble chart, and on the right, it shows the summary of the "assignment" aspect.

suggestions from student comments (Ramesh et al., 2014; Stupans et al., 2016; Gottipati et al., 2018; Unankard and Nadee, 2019; Hynninen et al., 2019). The common goal of these works is to answer some research questions (e.g., what are sentiment differences across courses and students). In contrast, we provide a demonstration tool of SET results to help instructors gain insights on their own and to allow others have access to organized summaries.

The most relevant works to ours are SUFAT (Pyasi et al., 2018) and Palaute (Grönberg et al., 2021) – two analytic tools for student comments. They both support sentiment analysis and LDA topic models (Blei et al., 2003), while we use more advanced RoBERTa-based sentiment analysis and weakly-supervised aspect extraction models. SUFAT requires users to install the tool and load SET files locally, while our online website directly reads the data from the SET instrument. More importantly, none of them conducts human evaluations, which makes it unclear if their tools are useful from the actual users' perspectives. Therefore, we develop the first demonstration system that uses an actual university SET instrument and is evaluated by university instructors who are interpreting their own evaluations.

3 SET Data

We use Student Evaluations of Teaching (SETs) data of over four academic terms (Fall 2017, Spring 2018, Fall 2018, Spring 2019) collected at UNC Chapel Hill. We utilize "semester + course number" as the specific identity of each course. We assume each course has just one instructor.[1] In total, there are about 5.6K courses and 298K SETs. Each SET is an evaluation questionnaire assigned to a student for a specific course they enrolled in, including both quantitative and open-ended questions.

UNC's SET instrument includes a series of evaluation questions assessing different aspects of the course and instructor. For demonstration, we select four representative questions – two quantitative and two open-ended items. For quantitative items, we choose *Overall, this course was excellent (Course Rate)* and *Overall, this instructor was an effective teacher (Instructor Rate)*, showing students' overall ratings on the course and instructor performance. Both items are based on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). For open-ended items, we choose *Comments on overall assessment of this course (Course Comments)* and *Comments on overall assessment of this instructor (Instructor Comments)*. Our system can be easily extended to the full set of SET questions.

Because completing the SET form is not mandatory, the average response rates of the two quantitative items we choose are 46% and 43% respectively, and even lower response rates are observed for the two open-ended items: 17% and 16%, respectively.

[1] This is not always true, and we will deal with co-teaching situations in the next version of our system.

4 System Description

After logging in, instructors can select to display their SET results of which semester and which course. The dashboard shows two main sections: *Rating Analysis* and *Comment Analysis*. See screenshots of our demo in Fig. 6 (SETSUM V1.0) and Fig. 7 (SETSUM V1.1) in Appendix.

4.1 Rating Analysis

For each of the two quantitative questions, we show the following statistics.

Response Rate. Since students do not always respond to every SET question, knowing how many students responded is critical for interpreting the generalizability and representativeness of the results. The standard report (Fig. 5) provides the number of responses for each question. To make this information stand out, we use a circular packing chart to describe the proportion of students who answered the question in comparison to the total enrollment of the course (Fig. 2).

Sentiment Distribution. The standard SET report summarizes quantitative ratings by mean, median, standard deviation, and percentages of the 5 rating options. Instead, in SETSUM V1.0, we simplify ratings to be either *positive* (4 and 5) or *negative* (3 or lower). We show the positive vs. negative ratings via a pie chart (Fig. 3 in Appendix). However, after conducting human evaluations on SETSUM V1.0, we received feedback from instructors preferring the original 5-point scale distribution. Thus, in SETSUM V1.1, we include a detailed breakdown of all scores.

4.2 Comment Analysis

For open-ended questions, besides the option to view all raw comments as the standard report (by clicking the “View Raw Comments” button in SETSUM V1.0 or the “Table View” button in SETSUM V1.1), we provide the following new features.

4.2.1 Basic Statistics

We present the **Response Rate** of open-ended questions also by a circular packing chart. Student comments are raw texts without sentiment labels. Therefore, we develop a sentiment analysis model (Sec. 4.2.2) and get the sentiment of each comment sentence. Then, we display the **Sentiment Distribution** (positive vs. negative ratio) via a pie chart for instructors to acquire an overview of students' sentiments expressed in comments.

4.2.2 Sentiment Analysis

As mentioned in Sec. 2, many existing works have conducted sentiment analysis on SET data (Wen et al., 2014; Azab et al., 2016; Baddam et al., 2019). In UNC's SET instrument, no sentiment labels are explicitly related to student comments. To train a sentiment analysis model, we pair *Course Comments* with the *Course Rate* since they are both overall assessment of the course. Similarly, we pair *Instructor Comments* with *Instructor Rate*.

We want to get sentence-level sentiments to compute the overall sentiment of each aspect (Sec. 4.2.3) and conduct summarization (Sec. 4.2.4). However, ratings are comment-level sentences. Thus, we first train a comment-level sentiment analysis model, and then we use it to predict the sentiments of each comment sentence.

4.2.3 Aspect Extraction

Students usually comment on some common *aspects* of the course, e.g., grade, assignment. Previous works resort to LDA (Blei et al., 2003) to learn *topics* from student comments (Ramesh et al., 2014; Pyasi et al., 2018; Grönberg et al., 2021). We argue that each topic learned from LDA is a set of words that is hard to be assigned a post hoc name, and topics sometimes lack distinctions (Ramesh et al., 2014). Therefore, we apply a weakly-supervised aspect extraction model, MATE (Angelidis and Lapata, 2018), that can extract aspects from comments using a set of pre-defined aspects.

MATE. *Multi Seed Aspect Extractor* (MATE) (Angelidis and Lapata, 2018) is derived from *Aspect-Based Autoencoder* (ABAE) (He et al., 2017). ABAE learns a sentence-level aspect predictor without supervision by reconstructing the sentence embedding as a linear combination of aspect embeddings. Assume $\mathbf{v}_s$ is the sentence embedding and $\mathbf{A}$ is a matrix of aspect embeddings, ABAE first predicts aspects: $\mathbf{p}_s^{aspect} = softmax(\mathbf{W}\mathbf{v}_s + \mathbf{b})$, and then reconstructs the sentence vector: $\mathbf{r}_s = \mathbf{A}^\top \mathbf{p}_s^{aspect}$. The objective is a max-margin loss using random sentences n_i as negative examples:

$$\mathcal{L} = \sum_s \sum_i max(0, 1 - \mathbf{r}_s\mathbf{v}_s + \mathbf{r}_s\mathbf{v}_{n_i})$$

Similar to LDA, ABAE has to interpret the learned aspects post hoc. To address this, MATE predefines a set of aspects by humans, and each aspect is given a set of *seed words*. Concatenating seed word embeddings together forms an aspect seed matrix $\mathbf{A}_i$, and the final aspect embedding matrix $\mathbf{A} = [\mathbf{A}_1^\top \mathbf{z}_1, ..., \mathbf{A}_K^\top \mathbf{z}_K]$, where $\mathbf{z}_i$ is a weight vector of seed words.

Aspect Annotation. To pre-label aspects of student comments and get seed words for each aspect, we randomly sample 100 comments for each of the two open-ended questions from the entire corpus and split them into sentences. Two human annotators (two authors) work together, attribute one or more aspects to each sentence, and label the corresponding aspect sentiments (positive or negative). Table 3 in Appendix shows two examples. In the end, we obtain 14 and 10 aspects of comments on course and instructor, respectively, and their terminology is defined in Table 5, 6 in Appendix. With the annotations, we calculate *clarity* scores (Cronen-Townsend et al., 2002) of each word w.r.t. each aspect (see details in Appendix A). The higher the clarity score, the more likely the word will appear in sentences of a specific aspect. We manually select 5 top-scored words for each aspect while removing noise (stopwords, names). Their scores are re-normalized to add up to 1. Table 4 shows the 5 seed words (plus weights) for each aspect.

Visualization. After training the MATE model, we predict the aspects of each comment sentence. We select all aspects that have $p_s^{aspect} > 0.4$. The threshold (0.4) is tuned on the subset with aspect annotations. Then, for each open-ended question of each course, we visualize its aspect distribution via a bubble chart (Fig. 1). Bubble size represents the number of sentences of this aspect. While bubble color denotes the aspect sentiment – the average of sentence-level sentiments, we chose accessible color palette (the more blue the more positive, the more yellow the more negative).

4.2.4 Extractive Summarization

After clustering comments by aspects, we want to provide a summary of each aspect. We first obtain the "centrality" of each sentence and then propose a method to extract summaries with high centrality, low redundancy, and balanced sentiments.

LexRank. For all the comment sentences under a certain aspect, we use LexRank (Erkan and Radev, 2004) to get the graph-based "centrality" of each sentence, where we use the cosine similarity of sentence embeddings from Sentence-BERT (Reimers and Gurevych, 2019). Intuitively, if a sentence is similar to many other sentences, it will be close to the "center" of the graph and thus it is prominent.

Algorithm 1: Summarization

```
Input: $S^a, K$
Output: $S'$
$S' \leftarrow \emptyset,\ S'_a \leftarrow S_a,\ k \leftarrow 1$;
while $k \le K$ do
    $s \leftarrow \arg\max_{s \in S'_a} J(s, S', S_a)$;
    $S' \leftarrow S' \cup \{s\}$;
    $S'_a \leftarrow S'_a - \{s\}$;
    $k \leftarrow k + 1$
end
```

Sentence Extraction Algorithm. Naively, we could extract the top central sentences as the summary. However, such summary sometimes includes redundant information and tends to only select positive sentences as they are more common. Inspired by Hsu and Tan (2021), we propose a greedy sentence extraction algorithm that optimizes three objectives on sentence selection: (1) maximizes centrality; (2) maximizes the difference between the sentence and other sentences extracted from previous steps; (3) minimizes the difference between the summary sentiment and the overall sentiment of the aspect. Algorithm 1 demonstrates our unsupervised extractive summarization algorithm, in which S_a represents all sentences under an aspect a, K is the number of sentences we want to extract (K=5), and S' is the target summary. Our learning objective (we want to maximize it) at each extraction step is written as:

$$J(s, S', S_a) = \text{centrality}_s - \text{cosine\_sim}(s, S') \\ - \text{senti\_diff}(S' \cup \{s\}, S_a)$$

Essentially, we want to extract a summary with high centrality, low redundancy, and a balanced sentiment. centrality_s is the centrality of sentence s. Following Hsu and Tan (2021), we define $\text{cosine\_sim}(s, S')$ as follows:

$$\text{cosine\_sim}(s, S') = \max_{s' \in S'} \text{cosine}(v_s, v_{s'})$$

where v_s and $v_{s'}$ are sentence embeddings from Sentence-BERT (Reimers and Gurevych, 2019). And we define $\text{senti\_diff}(S' \cup \{s\}, S_a)$, as the following:

$$\text{senti\_diff} = \left| \frac{\sum_{s' \in S' \cup \{s\}} p(s')}{|S' \cup \{s\}|} - \frac{\sum_{s' \in S_a} p(s')}{|S_a|} \right|$$

where p is the probability of positive sentiment predicted by our sentiment analysis model.

Visualization. Hovering any bubble in the aspect bubble chart will display its summary on the right (Fig. 1). Clicking on the aspect tab will display listed summary sentences within their the original comments to provide contextual information. A table of all sentences is on the bottom.

5 Evaluation & Results

5.1 Automatic Evaluation

Sentiment Analysis. We train two comment-level sentiment analysis models for *Course Comments* and *Instructor Comments* respectively. We split our data into training (90%) and development (10%) sets, and about 6.3K and 5.8K examples are in Course and Instructor development sets respectively. We first report comment-level sentiment prediction performance on the dev sets. Second, we use the comment-level sentiment analysis models to predict sentence-level sentiments during inference. To evaluate this, we use our aspect annotation data (Table 3 in Appendix), and we only use sentences with just one sentiment (i.e., all aspects are positive or negative), resulting in 202 and 230 testing examples for Course and Instructor. We report micro F1 (=accuracy) and macro F1. Table 1 shows the results. It can be seen that our models achieve reasonably good sentiment prediction performance, though they perform worse on predicting sentence-level sentiments than the comment level.

Sentiment Analysis	Course	Instructor
Comment-level micro F1	0.87	0.94
Comment-level macro F1	0.83	0.86
Sentence-level micro F1	0.83	0.90
Sentence-level macro F1	0.84	0.85

Table 1: Sentiment analysis results.

Summarization	Course		Instructor	
	Base.	Ours	Base.	Ours
Centrality↑	**1.13**	1.09	**1.14**	1.10
Redundancy↓	0.05	**0.03**	0.05	**0.02**
Sentiment Diff↓	0.41	**0.34**	0.43	**0.36**

Table 2: Summarization results.

Aspect Extraction. Similarly, we also train two aspect extraction models for *Course Comments* and *Instructor Comments* separately. We evaluate their performance by comparing to human annotated aspects using F1 score. In total, we have 213 and 234 testing examples for course and instructor models, and the average number of aspects is 1.38 and 1.31, respectively. We achieve F1 score of 48.6 for the course model and 48.9 for the instructor model, which are similar to the results of the MATE paper (Angelidis and Lapata, 2018). We also explore another approach by treating aspect extraction as a multi-label aspect classification task. We use half of the annotated data to finetune a RoBERTa-base (Liu et al., 2019) model and test on the other half annotated aspects. Our experiment shows improved F1 scores of 62.6 for the course model and 64.9 for the instructor model. We plan to combine RoBERTa and MATE to deploy a weakly-supervised RoBERTa-based MATE in our next version of website.

Summarization. Due to the lack of gold summaries, we use three metrics (*Centrality*, *Redundancy*, and *Sentiment Difference*) to evaluate our summarization approach and compare it to the baseline of extracting the top 5 central sentences. Please refer to Appendix C for detailed definitions of these three metircs. We randomly sampled 100 courses as the testing set to report the performance. Table 2 shows the results. As expected, our method leads to lower redundancy and sentiment difference than the baseline, though it scarifies some centrality.

5.2 Human Evaluation

It is critical to evaluate how our demonstration system is perceived by its primary users: instructors.

5.2.1 Evaluation Setup

Design a Survey. We design an evaluation survey using Qualtrics. Our complete survey can be found at SETSum Github Repo. In the survey, we first introduce the background and purpose. We define the standard PDF report *Usual Approach* and our SETSUM V1.0 as *Comparison Approach*, and then we ask instructors to compare the two approaches. The main survey body contains 5 parts of questions:

(1) *Rate the Usual Approach*: Without comparing to SETSUM, we ask how they rate the usefulness of standard SET reports in a 5-point scale: not at all, slightly, moderately, very, or extremely useful;

(2) *Rate* SETSUM *(Rating Analysis)*: Compared to the usual approach, instructors rate our new features of summarizing ratings in a slightly different 5-point scale: not at all useful, not useful, equally useful, useful, or very useful;

(3) *Rate* SETSUM *(Comments Analysis)*: Compared to the usual approach, we ask how useful each of our new features of summarizing comments is (using the same response anchors as (2)).

(4) *Rate the overall experience with* SETSUM: We ask if our website helps them interpret SET results more efficiently and/or with less bias (definitely not, probably not, might or might not, probably yes, definitely yes) as well as if they prefer the standard SET report or our website or both.

(5) *Comments*: Instructors may leave additional comments on the website under development.

Invite Instructors. We invited 15 professors at UNC, who taught large introductory courses within the studied period (4 semesters). We estimated the survey to take 20-30 minutes, and each participant was offered a $25 gift card to a campus coffee shop. In the end, 10 instructors from 8 different departments completed the survey successfully.

5.2.2 Results Analysis

Fig. 4 shows the survey results, and the Qualtrics report can be found at SETSum Github Repo. Here, we summarize some main takeaways.

Instructors have positive opinions about the standard SET report. 8 out of 10 (and 6 out of 10) instructors think the PDF report is *moderately to extremely useful* in summarizing students' ratings (and comments), respectively. This demonstrates the well-perceived usefulness of existing SET reports by instructors, though they are less satisfied with the comment summarization.

New features introduced on SETSUM are perceived to be useful or very useful. On average, for rating analysis, 7 out of 10 instructors think each of the 2 new features (response rate and sentiment distribution) is *useful* or *very useful*, and for comments analysis, 8.8 out of 10 instructors on avg. think each of the 5 new features (response rate, sentiment distribution, topic bubbles, summary sentences, showing original comments for each summary sentence) is *useful* or *very useful*, while fewer instructors (5.5 out of 10 on avg.) think the scatter plot[2] and the table showing all comment sentences are *useful* or *very useful*. Overall, most instructors perceive our SETSUM as being useful.

SETSUM helps all instructors interpret SET results more efficiently, and it helps some instructors interpret SET results with less bias. All instructors agree that SETSUM helps them interpret SETs *more efficiently* (i.e., probably to definitely yes). 4 out of 10 instructors think it helps them understand SETs *with less bias*.

Instructors prefer SETSUM than the standard report or would like to have both. Lastly, 6 out of 10 instructors prefer SETSUM compared to the usual approach, while 4 instructors would like to have both approaches.

Constructive suggestions. We identify the following suggestions from instructors' comments for improving our future version: (1) The accuracy of the sentiment analysis and aspect extraction models can still be improved. (2) Many instructors prefer the complete display of ratings in the 5-point scale, rather than presenting only a positive v.s. negative ratio. (3) Instructors without a computer science background had difficulty understanding concepts like "centrality". So far, we addressed (2) and (3) in SETSUM V1.1 by providing the 5-point scale rating distribution and adding detailed explanations for each Machine Learning related modules.

Overall, instructors show a very positive attitude towards our SETSUM demonstration system and provided important suggestions and direction for our future work.

6 Conclusion

In this work, we propose SETSUM, a system to summarize and visualize results from student evaluations of teaching. We integrate NLP, statistical, visualization, and web service techniques. We are among the few researchers to build a tool for instructor use and are the first to evaluate the tool by university professors. Our results demonstrate that our system is promising at improving the SET report paradigm and helps instructors gain insights from their SETs more efficiently. In the future, we will keep improving the sentiment analysis and aspect extraction models to provide more accurate summarization of SET results. The instructor evaluation offered key recommendations for the next iterations of the system. We will incorporate more functions to our system, including allowing instructors to compare different courses and track their own teaching history of their courses as well as developing a separate administrator dashboard to identify themes across academic courses, departments, and programs.

[2] We had a scatter plot showing all comment sentences in SETSUM V1.0, which was removed from SETSUM V1.1.

7 Ethical Considerations

As mentioned earlier, SETs have multiple functions such as (1) faculty members examining their teaching performance in a diagnostic way, (2) allowing institution leaders to review and describe the quality of course offerings, (3) part of an instructor's larger portfolio to demonstrate their teaching history during high-stakes settings, and (4) summaries being released to students to guide them with course selections. Given this wide range of uses for the SET summaries, our work's purpose is to take initial steps towards developing thoughtful, accurate, and well-designed representations that can be provided to draw accurate inferences about teaching quality, course design, and student learning. However, it is also critical to examine all aspects of SETs through an ethical lens. Errors in NLP-based analysis could lead to misinterpretation and inaccurate judgments in high-stakes settings. Therefore in the following, we discuss how each module of our SETSum website affects the interpretation of SET results.

For quantitative items, we provide visualizations of two statistics that are directly computed from SET data. Therefore, no errors or biases should be introduced compared to the standard PDF report. In fact, some instructors who participated in our evaluations say that the *response rate* feature for each individual question helps them understand the results with less bias.

For open-ended items, to obtain their sentiment distributions, we develop sentiment analysis models to obtain sentence-level sentiments. Though we obtain good sentiment prediction performance (Table 1), errors are inevitable. We use these features to demonstrate the relative number of positive to negative comments (ratio). In general, unless very few students evaluate a course (low response rate for comments), the system can still convey the information fairly well. Another important feature that we develop as part of this system is to group comment sentences by aspects. Although we achieve similar aspect prediction F1 scores consistent with past research, we find that the results are not precise enough yet for widespread use. Our human evaluators notice that some sentences from the open-ended comments are inaccurately clustered. Therefore, in future iterations of this system, we believe it is very important to develop a more accurate aspect extraction model. The final important feature is the unsupervised extraction summarization module. We choose an extraction method because it does not suffer from faithfulness (not staying true to the source) issues as abstractive methods (Cao et al., 2018). Meanwhile, our algorithm extracts summaries with more balanced sentiments (Table 2). Nonetheless, we hope to find a summarization approach that aligns more closely with the sentiments underlying the students' comments.

Though instructors express positive attitudes towards our system and 4 instructors think it help them understand SETs with less bias, we believe that additional thorough evaluations need to be conducted in the future. Outside of SETs, our work recognizes the different ways, reporters, and methods that could be used to assess teaching effectiveness, including but not limited to peer reports, analysis of classroom sound, student learning, and an instructor's own teaching portfolio.

Finally, at this time our system is designed to be used and reviewed by instructors or other reviewers, and it *does not* directly make any broad judgments or decisions (e.g., whether the instructor is qualified for promotion). The primary end-users of the system should be instructors who wish to analyze their SET findings more thoroughly and acquire the main takeaways more efficiently. Other reviewers and administrators can use the system to view the SET findings in a broader scope, such as reading the report summary per department or per division. Overall, the goal of SETSUM is to help instructors and other reviewers to understand more of students' needs and make improvements to future course design.

Acknowledgments

We thank the reviewers for their helpful comments. We thank the UNC instructors who participated in our human evaluations. We would also like to thank Heather Thompson from the Office of Undergraduate Curricula for providing the data and Rob Ricks from the Office of Institutional Research and Assessment for additional data preparation activities. This work was supported by NSF-CAREER Award 1846185, NSF-AI Engage Institute DRL-2112635, a Bloomberg Data Science Ph.D. Fellowship, and the Howard Hughes Medical Institute Inclusive Excellence 3 Grant.

References

Fadia Nasser-Abu Alhija and Barbara Fresko. 2009. Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational evaluation*, 35(1):37–44.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Mahmoud Azab, Rada Mihalcea, and Jacob Abernethy. 2016. Analysing ratemyprofessors evaluations across institutions, disciplines, and cultures: The tell-tale signs of a good professor. In *International Conference on Social Informatics*, pages 438–453. Springer.

Swathi Baddam, Prasad Bingi, and Syed Shuva. 2019. Student evaluation of teaching in business education: Discovering student sentiments using text mining techniques. *e-Journal of Business Education and Scholarship of Teaching*, 13(3):1–13.

Esenc M Balam and David M Shannon. 2010. Student ratings of college teaching: A comparison of faculty and their students. *Assessment & Evaluation in Higher Education*, 35(2):209–221.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yining Chen and Leon B Hoshower. 2003. Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & evaluation in higher education*, 28(1):71–88.

Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306.

Sam Cunningham-Nelson, Mahsa Baktashmotlagh, and Wageeh Boles. 2018. Visually exploring sentiment and keywords for analysing student satisfaction data. *Proceedings of the 29th Australasian Association of Engineering Education (AAEE 2018)*, pages 1–7.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Swapna Gottipati, Venky Shankararaman, and Jeff Rongsheng Lin. 2018. Text analytics approach to extract course improvement suggestions from students' feedback. *Research and Practice in Technology Enhanced Learning*, 13(1):1–19.

Bryan W Griffin. 2001. Instructor reputation and student ratings of instruction. *Contemporary educational psychology*, 26(4):534–552.

Niku Grönberg, Antti Knutas, Timo Hynninen, and Maija Hujala. 2021. Palaute: An online text mining tool for analyzing written student course feedback. *IEEE Access*, 9:134518–134529.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397.

Khe Foon Hew, Xiang Hu, Chen Qiao, and Ying Tang. 2020. What predicts student satisfaction with moocs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, 145:103724.

Chao-Chun Hsu and Chenhao Tan. 2021. Decision-focused summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 117–132.

Maija Hujala, Antti Knutas, Timo Hynninen, and Heli Arminen. 2020. Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & Education*, 157:103965.

Timo Hynninen, Antti Knutas, Maija Hujala, and Heli Arminen. 2019. Distinguishing the themes emerging from masses of open student feedback. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 557–561. IEEE.

David E Kanouse and L Reid Hanson Jr. 1987. Negativity in evaluations. In *Preparation of this paper grew out of a workshop on attribution theory held at University of California, Los Angeles, Aug 1969.* Lawrence Erlbaum Associates, Inc.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

James A Kulik. 2001. Student ratings: Validity, utility, and controversy. *New directions for institutional research*, 2001(109):9–25.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Pablo Marshall. 2021. Contribution of open-ended questions in student evaluation of teaching. *Higher Education Research & Development*, pages 1–14.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Angela R Penny and Robert Coe. 2004. Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of educational research*, 74(2):215–253.

Siddhant Pyasi, Swapna Gottipati, and Venky Shankararaman. 2018. Sufat-an analytics tool for gaining insights from student feedback comments. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, and Lise Getoor. 2014. Understanding mooc discussion forums using seeded lda. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 28–33.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Daniel Febrian Sengkey, Agustinus Jacobus, and Fabian Johanes Manoppo. 2019. Implementing support vector machine sentiment analysis to students' opinion toward lecturer in an indonesian public university. *Journal of Sustainable Engineering: Proceedings Series*, 1(2):194–198.

Penny M Simpson and Judy A Siguaw. 2000. Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, 22(3):199–213.

Pieter Spooren, Bert Brockx, and Dimitri Mortelmans. 2013. On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4):598–642.

Ieva Stupans, Therese McGuren, and Anna Marie Babey. 2016. Student evaluation of teaching: A study exploring student rating instrument free-form text comments. *Innovative Higher Education*, 41(1):33–42.

Sayan Unankard and Wanvimol Nadee. 2019. Topic detection for online course feedback using lda. In *International Symposium on Emerging Technologies for Education*, pages 133–142. Springer.

Miaomiao Wen, Diyi Yang, and Carolyn Rose. 2014. Sentiment analysis in mooc discussion forums: What does it tell us? In *Educational data mining 2014*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

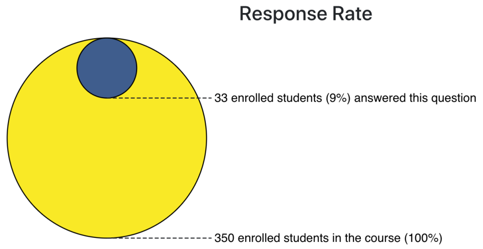


Figure 2: A circular packing chart describes the response rate.

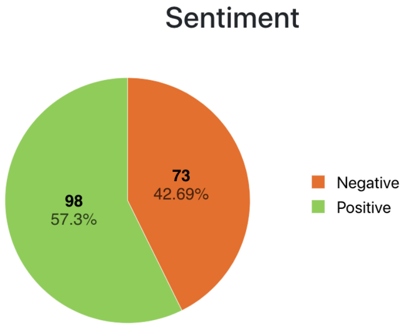


Figure 3: A pie chart describes the sentiment distribution.

Francisco Zabaleta. 2007. The use and misuse of student evaluations of teaching. *Teaching in higher education*, 12(1):55–76.

Appendix

A Clarity Score

To identify seed words for each aspect. We compute the clarity score (Cronen-Townsend et al., 2002; Angelidis and Lapata, 2018) of each word with respect to each aspect. The score measures how likely it is to observe word w in comments of aspect a: $score_a(w) = t_a(w) log \frac{t_a(w)}{t(w)}$, where $t_a(w)$ is the tf-idf score of w in comments of aspect a and $t(w)$ is that in all comments.

B Implementation Details

Sentiment Analysis. We finetuned a RoBERT-large model (Liu et al., 2019) using HuggingFace's Transformers (Wolf et al., 2020) for 5 epochs and chose the best performed checkpoint on the development set. We used the AdamW optimizer

(Loshchilov and Hutter, 2018) with learning rate 1e-5 and batch size 16.

Aspect Extraction. We used NLTK to conduct sentence and word segmentation. We initialized the MATE model using GloVe embeddings (Pennington et al., 2014). During the training, the word embeddings, seed word matrices, and seed weight vectors were fixed and we trained the model for 10 epochs using Adam optimizer (Kingma and Ba, 2015) with learning rate 10^{-1} and batch size 50. We also experimented the multi-label classification approach by finetuning a RoBERTa-base model (Liu et al., 2019) for 10 epochs using AdamW optimizer (Loshchilov and Hutter, 2018) with learning rate 2e-5 and batch size 16.

Website. We developed the website using the React framework for the front-end interface. For the back-end, we set up a database with Firebase and created a RESTful API with Firebase Cloud Functions. Our website is deployed to Netlify.com for an online demonstration.

C Summarization Evaluation Metrics

We define the *Centrality* metric as the average centrality of summary sentences. The higher the metric is, the better. Assume the summary to evaluate is S'.

$$\text{Centrality}(S') = \frac{\sum_{s \in S'} \text{centrality}_s}{|S'|}$$

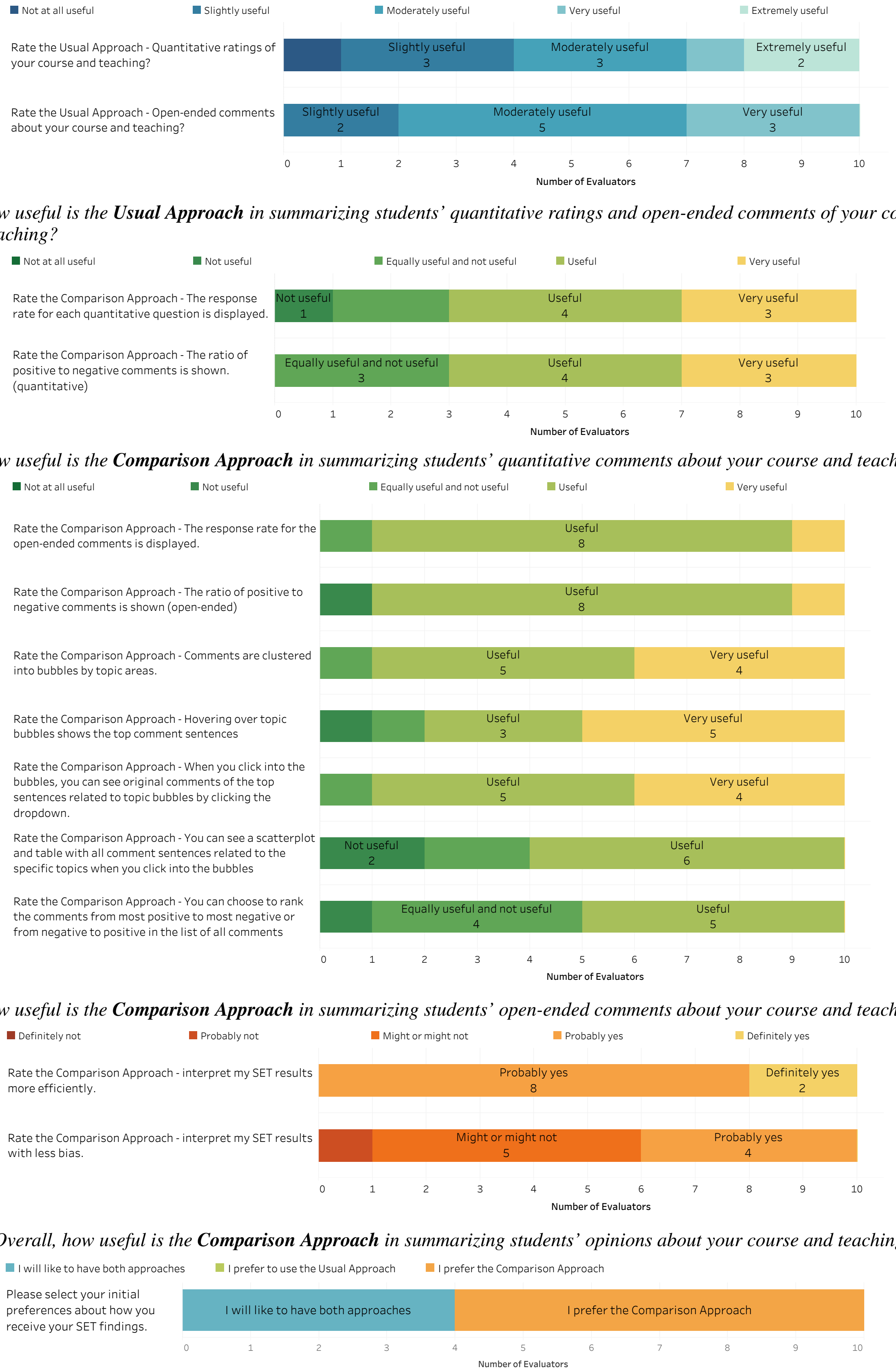
We compute the information *Redundancy* within a summary S' by taking the average of cosine similarity among sentences. We use sentence embeddings from Sentence-BERT (Reimers and Gurevych, 2019) to compute cosine similarities. The lower the metric is, the better.

$$\text{Redun}(S') = \frac{\sum_{s \in S'} \max\limits_{s' \in S' - \{s\}} \text{cosine}(v_s, v_{s'})}{|S'|}$$

We first compute the average sentiments for the summary S' and all sentences under the aspect S_a, respectively. Then, we take their absolute difference as the final score of *Sentiment Difference*. The lower the metric is, the better.

$$\text{Senti\_diff} = |\frac{\sum_{s \in S'} p(s)}{|S'|} - \frac{\sum_{s \in S_a} p(s)}{|S_a|}|$$

where p is the probability of positive sentiment predicted by our sentiment analysis model.



(a) *How useful is the* ***Usual Approach*** *in summarizing students' quantitative ratings and open-ended comments of your course and teaching?*

(b) *How useful is the* ***Comparison Approach*** *in summarizing students' quantitative comments about your course and teaching?*

(c) *How useful is the* ***Comparison Approach*** *in summarizing students' open-ended comments about your course and teaching?*

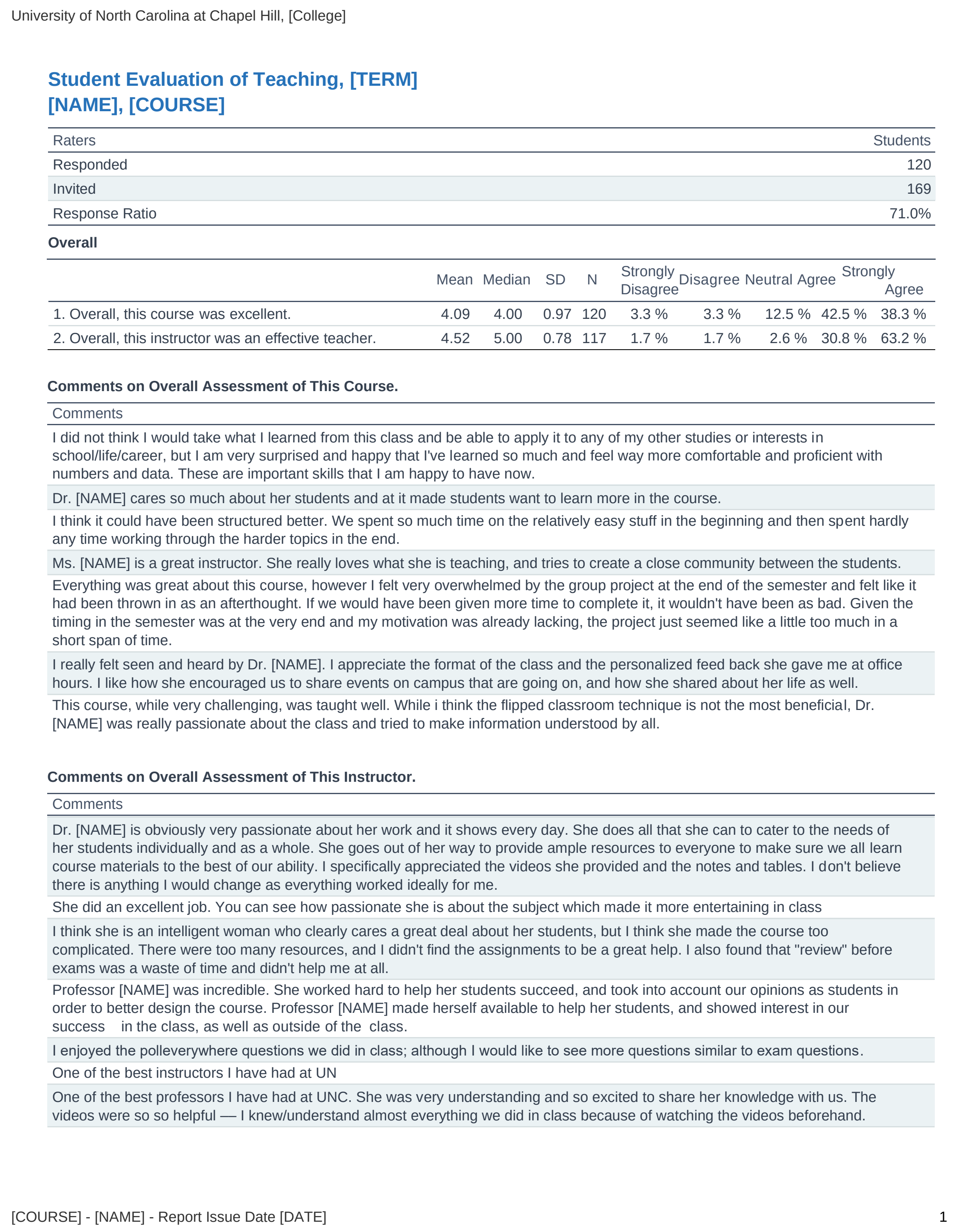
(d) *Overall, how useful is the* ***Comparison Approach*** *in summarizing students' opinions about your course and teaching?*

(e) *Overall, please select your initial preference about how you receive your SET findings.*

Figure 4: Results of human evaluation comparing the standard SET report (the Usual Approach) and the SETSUM v1.0 website (the Comparison Approach).

SET Question	Comment Sentence	(Aspect, Sentiment)
Comments on overall assessment of this course	Because, even though the lecture was fine the exams were brutal or was just wrong because of the answer key being wrong.	(content, positive); (exam, negative)
Comments on overall assessment of this instructor	The instructor was clear at explaining information and fairly evaluating all assignments.	(delivery, positive); (grade, positive)

Table 3: Two examples of Aspect Annotation.



University of North Carolina at Chapel Hill, [College]

Student Evaluation of Teaching, [TERM]
[NAME], [COURSE]

Raters	Students
Responded	120
Invited	169
Response Ratio	71.0%

Overall

	Mean	Median	SD	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. Overall, this course was excellent.	4.09	4.00	0.97	120	3.3 %	3.3 %	12.5 %	42.5 %	38.3 %
2. Overall, this instructor was an effective teacher.	4.52	5.00	0.78	117	1.7 %	1.7 %	2.6 %	30.8 %	63.2 %

Comments on Overall Assessment of This Course.

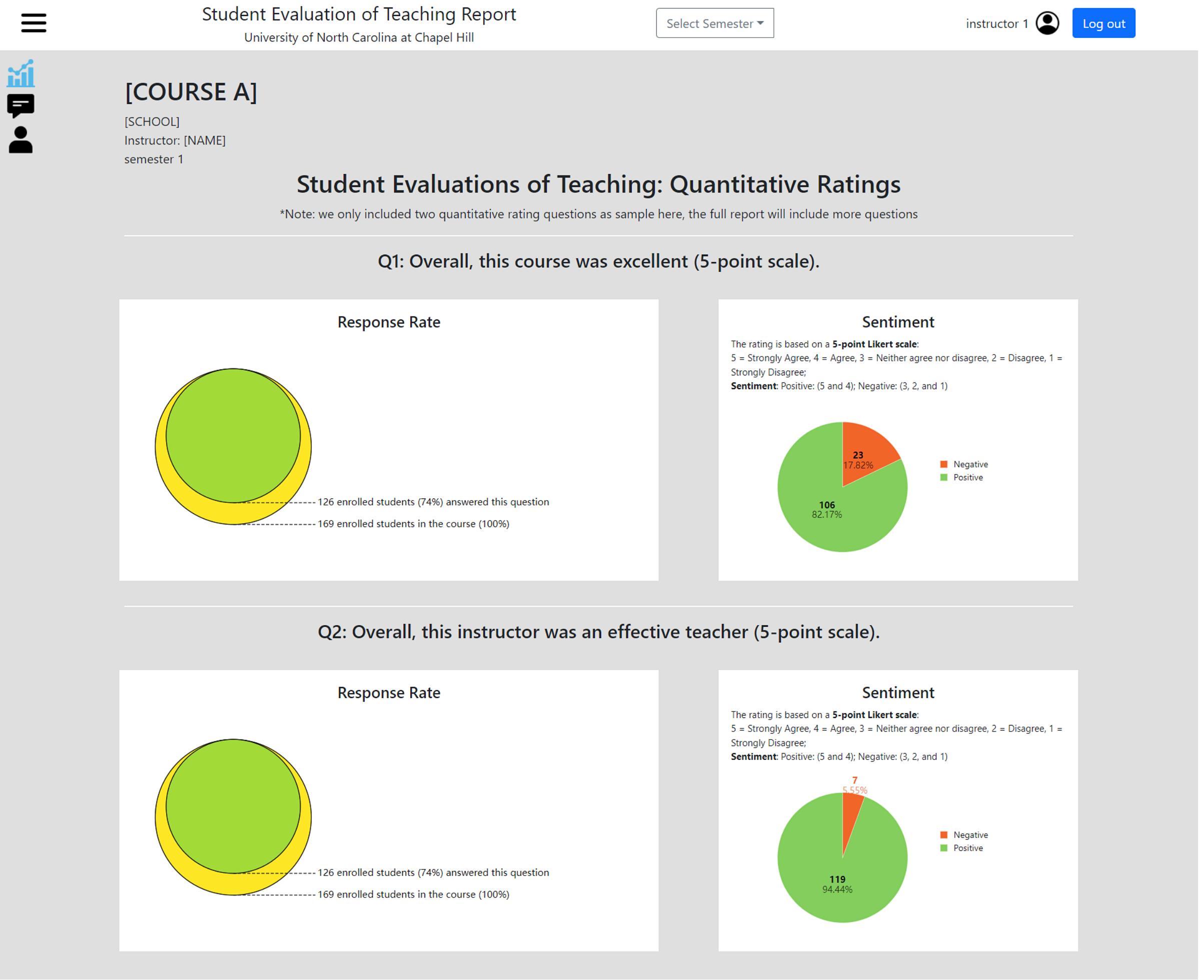
Comments
I did not think I would take what I learned from this class and be able to apply it to any of my other studies or interests in school/life/career, but I am very surprised and happy that I've learned so much and feel way more comfortable and proficient with numbers and data. These are important skills that I am happy to have now.
Dr. [NAME] cares so much about her students and at it made students want to learn more in the course.
I think it could have been structured better. We spent so much time on the relatively easy stuff in the beginning and then spent hardly any time working through the harder topics in the end.
Ms. [NAME] is a great instructor. She really loves what she is teaching, and tries to create a close community between the students.
Everything was great about this course, however I felt very overwhelmed by the group project at the end of the semester and felt like it had been thrown in as an afterthought. If we would have been given more time to complete it, it wouldn't have been as bad. Given the timing in the semester was at the very end and my motivation was already lacking, the project just seemed like a little too much in a short span of time.
I really felt seen and heard by Dr. [NAME]. I appreciate the format of the class and the personalized feed back she gave me at office hours. I like how she encouraged us to share events on campus that are going on, and how she shared about her life as well.
This course, while very challenging, was taught well. While i think the flipped classroom technique is not the most beneficial, Dr. [NAME] was really passionate about the class and tried to make information understood by all.

Comments on Overall Assessment of This Instructor.

Comments
Dr. [NAME] is obviously very passionate about her work and it shows every day. She does all that she can to cater to the needs of her students individually and as a whole. She goes out of her way to provide ample resources to everyone to make sure we all learn course materials to the best of our ability. I specifically appreciated the videos she provided and the notes and tables. I don't believe there is anything I would change as everything worked ideally for me.
She did an excellent job. You can see how passionate she is about the subject which made it more entertaining in class
I think she is an intelligent woman who clearly cares a great deal about her students, but I think she made the course too complicated. There were too many resources, and I didn't find the assignments to be a great help. I also found that "review" before exams was a waste of time and didn't help me at all.
Professor [NAME] was incredible. She worked hard to help her students succeed, and took into account our opinions as students in order to better design the course. Professor [NAME] made herself available to help her students, and showed interest in our success in the class, as well as outside of the class.
I enjoyed the polleverywhere questions we did in class; although I would like to see more questions similar to exam questions.
One of the best instructors I have had at UN
One of the best professors I have had at UNC. She was very understanding and so excited to share her knowledge with us. The videos were so so helpful — I knew/understand almost everything we did in class because of watching the videos beforehand.

[COURSE] - [NAME] - Report Issue Date [DATE] 1

Figure 5: The standard PDF SET report.



Student Evaluation of Teaching Report
University of North Carolina at Chapel Hill
Select Semester
instructor 1
Log out

[COURSE A]

[SCHOOL]
Instructor: [NAME]
semester 1

Student Evaluations of Teaching: Quantitative Ratings

*Note: we only included two quantitative rating questions as sample here, the full report will include more questions

Q1: Overall, this course was excellent (5-point scale).

Response Rate

126 enrolled students (74%) answered this question
169 enrolled students in the course (100%)

Sentiment

The rating is based on a **5-point Likert scale**:
5 = Strongly Agree, 4 = Agree, 3 = Neither agree nor disagree, 2 = Disagree, 1 = Strongly Disagree;
Sentiment: Positive: (5 and 4); Negative: (3, 2, and 1)

Q2: Overall, this instructor was an effective teacher (5-point scale).

Response Rate

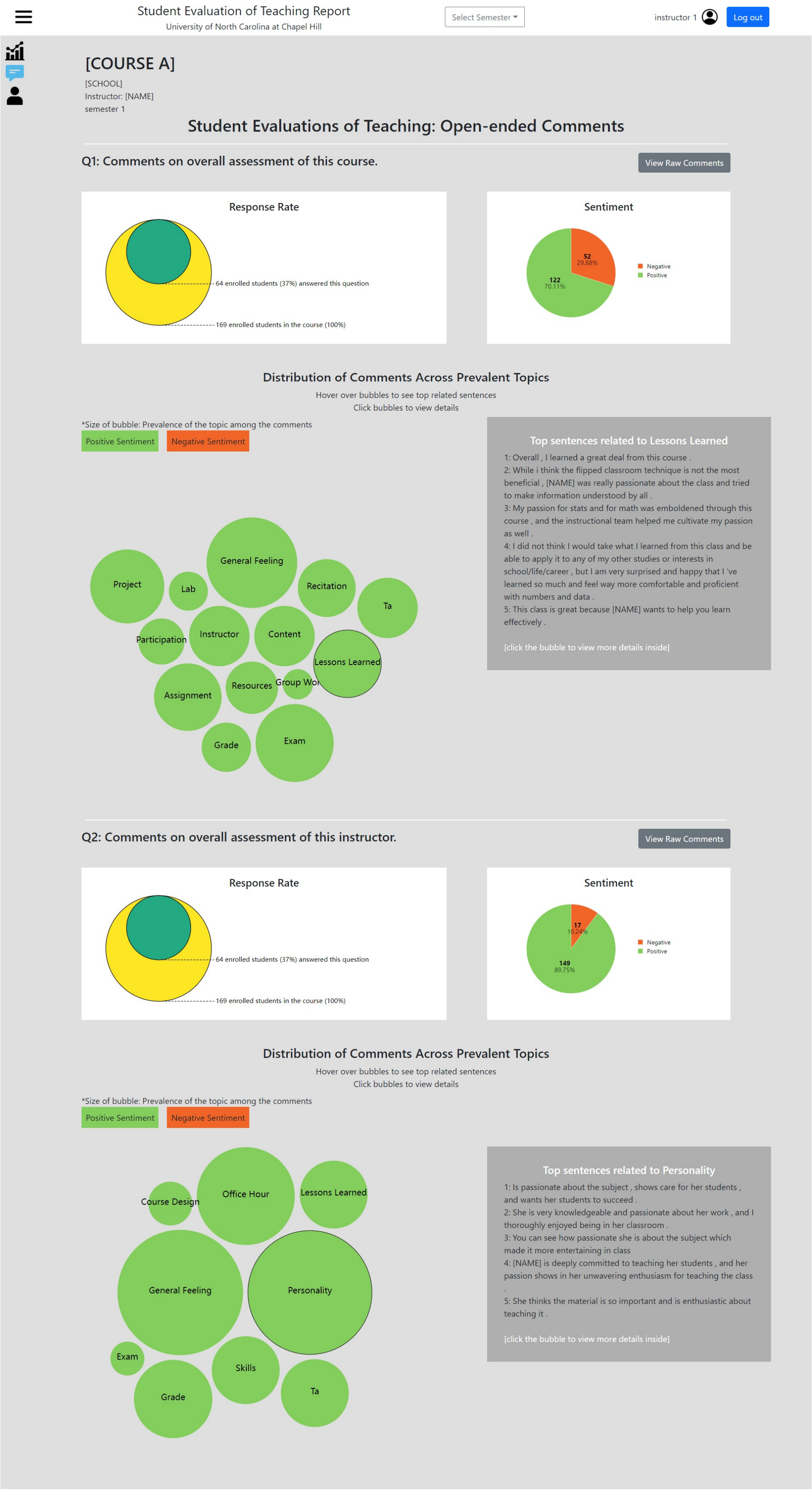
126 enrolled students (74%) answered this question
169 enrolled students in the course (100%)

Sentiment

The rating is based on a **5-point Likert scale**:
5 = Strongly Agree, 4 = Agree, 3 = Neither agree nor disagree, 2 = Disagree, 1 = Strongly Disagree;
Sentiment: Positive: (5 and 4); Negative: (3, 2, and 1)

(a) A page shows the Rating Analysis (Quantitative Questions) part.

Figure 6: Screenshots of SETSUM v1.0 (Part1, see Part2 in the next page).



[COURSE A]

[SCHOOL]
Instructor: [NAME]
semester 1

Student Evaluations of Teaching: Open-ended Comments

Q1: Comments on overall assessment of this course.

Distribution of Comments Across Prevalent Topics

Hover over bubbles to see top related sentences
Click bubbles to view details

*Size of bubble: Prevalence of the topic among the comments

Top sentences related to Lessons Learned

1: Overall , I learned a great deal from this course .
2: While i think the flipped classroom technique is not the most beneficial , [NAME] was really passionate about the class and tried to make information understood by all .
3: My passion for stats and for math was emboldened through this course , and the instructional team helped me cultivate my passion as well .
4: I did not think I would take what I learned from this class and be able to apply it to any of my other studies or interests in school/life/career , but I am very surprised and happy that I 've learned so much and feel way more comfortable and proficient with numbers and data .
5: This class is great because [NAME] wants to help you learn effectively .

[click the bubble to view more details inside]

Q2: Comments on overall assessment of this instructor.

Distribution of Comments Across Prevalent Topics

Hover over bubbles to see top related sentences
Click bubbles to view details

*Size of bubble: Prevalence of the topic among the comments

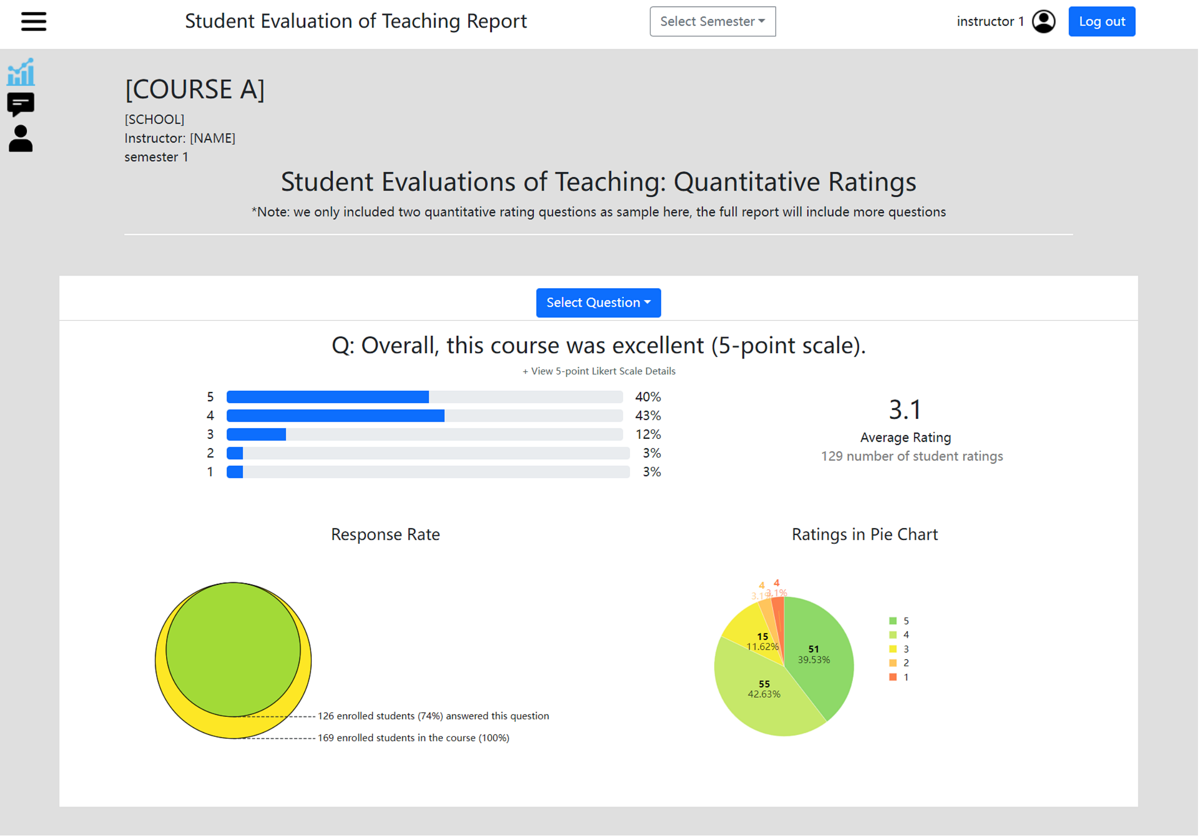
Top sentences related to Personality

1: Is passionate about the subject , shows care for her students , and wants her students to succeed .
2: She is very knowledgeable and passionate about her work , and I thoroughly enjoyed being in her classroom .
3: You can see how passionate she is about the subject which made it more entertaining in class
4: [NAME] is deeply committed to teaching her students , and her passion shows in her unwavering enthusiasm for teaching the class .
5: She thinks the material is so important and is enthusiastic about teaching it .

[click the bubble to view more details inside]

(b) A page shows the Comments Analysis (Open-ended Questions) part.

Figure 6: Screenshots of SETSUM v1.0 (Part2).



Student Evaluation of Teaching Report
Select Semester ▾
instructor 1
Log out

[COURSE A]

[SCHOOL]
Instructor: [NAME]
semester 1

Student Evaluations of Teaching: Quantitative Ratings

*Note: we only included two quantitative rating questions as sample here, the full report will include more questions

Q: Overall, this course was excellent (5-point scale).

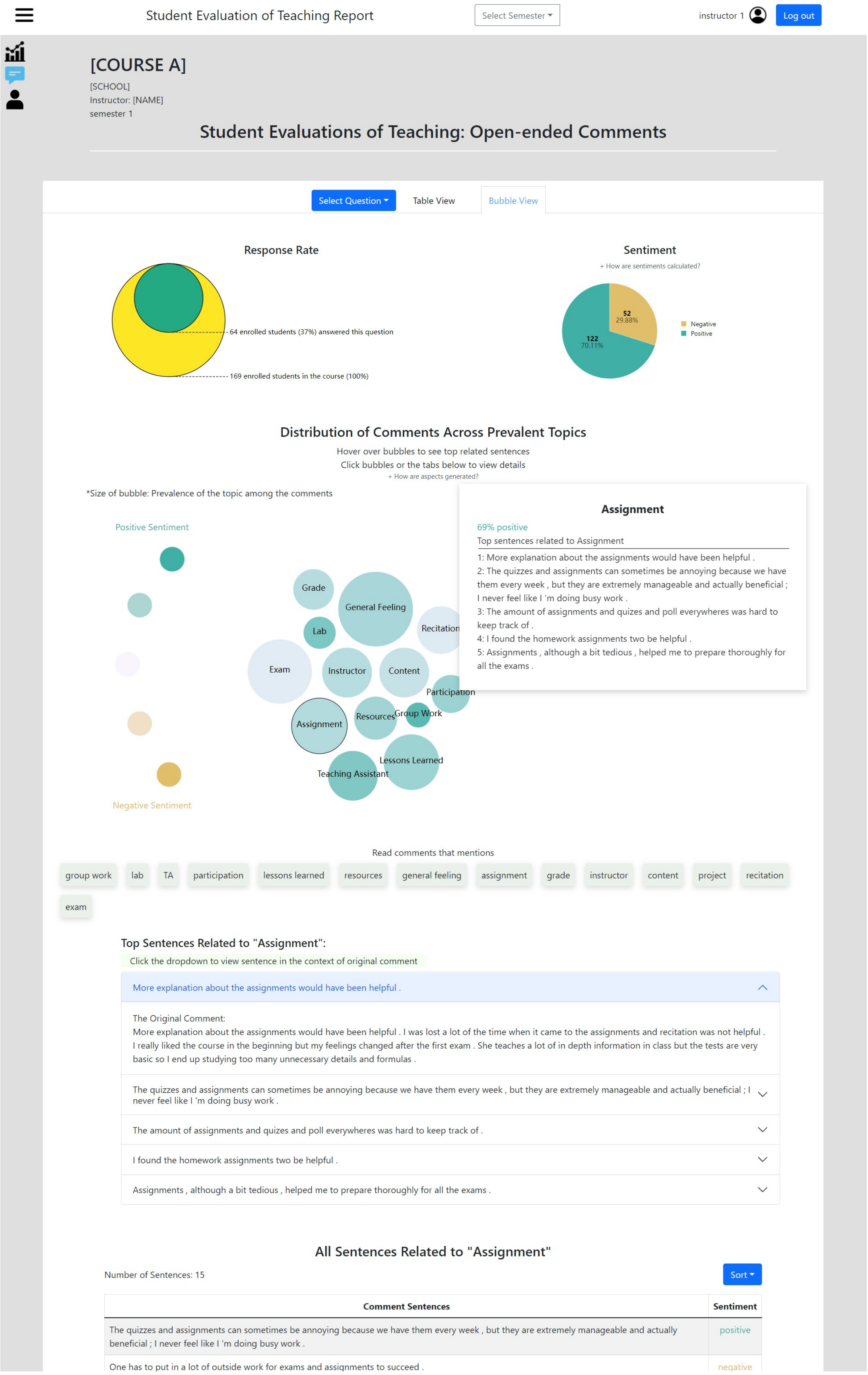
+ View 5-point Likert Scale Details
5
4
3
2
1
40%
43%
12%
3%
3%

3.1
Average Rating
129 number of student ratings

Response Rate
126 enrolled students (74%) answered this question
169 enrolled students in the course (100%)
Ratings in Pie Chart
4
3.1%
4
3.1%
15
11.62%
51
39.53%
55
42.63%
5
4
3
2
1

(a) A page shows the Rating Analysis (Quantitative Questions) part.

Figure 7: Screenshots of SETSUM v1.1 (Part1, see Part2 in the next page).



[COURSE A]

[SCHOOL]
Instructor: [NAME]
semester 1

Student Evaluations of Teaching: Open-ended Comments

Distribution of Comments Across Prevalent Topics

Hover over bubbles to see top related sentences
Click bubbles or the tabs below to view details
+ How are aspects generated?

*Size of bubble: Prevalence of the topic among the comments

Assignment

69% positive

Top sentences related to Assignment

1: More explanation about the assignments would have been helpful .
2: The quizzes and assignments can sometimes be annoying because we have them every week , but they are extremely manageable and actually beneficial ; I never feel like I 'm doing busy work .
3: The amount of assignments and quizes and poll everywheres was hard to keep track of .
4: I found the homework assignments two be helpful .
5: Assignments , although a bit tedious , helped me to prepare thoroughly for all the exams .

Read comments that mentions

Top Sentences Related to "Assignment":

Click the dropdown to view sentence in the context of original comment

More explanation about the assignments would have been helpful .

The Original Comment:
More explanation about the assignments would have been helpful . I was lost a lot of the time when it came to the assignments and recitation was not helpful . I really liked the course in the beginning but my feelings changed after the first exam . She teaches a lot of in depth information in class but the tests are very basic so I end up studying too many unnecessary details and formulas .

The quizzes and assignments can sometimes be annoying because we have them every week , but they are extremely manageable and actually beneficial ; I never feel like I 'm doing busy work .

The amount of assignments and quizes and poll everywheres was hard to keep track of .

I found the homework assignments two be helpful .

Assignments , although a bit tedious , helped me to prepare thoroughly for all the exams .

All Sentences Related to "Assignment"

Number of Sentences: 15

Comment Sentences	Sentiment
The quizzes and assignments can sometimes be annoying because we have them every week , but they are extremely manageable and actually beneficial ; I never feel like I 'm doing busy work .	positive
One has to put in a lot of outside work for exams and assignments to succeed .	negative

(b) A page shows the Comments Analysis (Open-ended Questions) part.

Figure 7: Screenshots of SETSUM v1.1 (Part2).

Aspect	Top words (normalized weight)
assignment	assignment (0.33), homework (0.31), concept (0.17), reading (0.13), exercise (0.07)
content	material (0.42), lecture (0.17), reading (0.15), subject (0.13), content (0.13)
course design	syllabus (0.21), requirement (0.20), communicated (0.20), wish (0.20), discussion (0.19)
exam	exam (0.31), test (0.24), question (0.20), answer (0.13), problem (0.13),
general feeling	course (0.33), enjoyed (0.25), favorite (0.19), challenging (0.12), hard (0.11)
grade	grading (0.39), feedback (0.19), harsh (0.16), midterm (0.16), easy (0.11)
group work	group (0.30), project (0.24), recitation (0.20), work (0.06), team (0.20)
instructor	professor (0.50), instructor (0.21), passionate (0.11), teach (0.11), condescending (0.06)
lab	lab (0.32), hand (0.20), report(0.20), grading (0.10), experiment (0.18)
lessons learned	learned (0.23), real (0.22), life (0.22), skill (0.19), understanding (0.14)
participation	discussion(0.30), speak(0.23), comfortable (0.18), participation (0.16), stressful (0.13)
project	project (0.30), instance (0.23), expectation (0.18), clearly (0.16), explained (0.13)
recitation	recitation (0.57), content (0.18), project (0.10), review (0.09), group (0.05)
resources	peer (0.20), mentor (0.20), book (0.20), software (0.20), reference (0.20)
teaching assistant (TA)	TA (0.42), job (0.20), helping (0.12), explained (0.06). available (0.20)

(a) Highest ranked words list for each aspect of *Comments on overall assessment of this course*.

Aspect	Top words (normalized weight)
course design	lecture (0.28), assignment (0.21), topic (0.18), activity (0.17), structured (0.17)
delivery	engaged (0.26), clear (0.22), lecture (0.22), example (0.16), explain (0.14)
exam	unfair (0.25), fair (0.25), exam (0.23), guide (0.20), question (0.08)
general feeling	professor (0.37), great (0.27), instructor (0.25), bad (0.05), overall (0.05)
grade	grade (0.36), passing (0.20), average (0.20), exam (0.13), comment (0.11)
lessons learned	conceptual (0.27), intellectual (0.27), learned (0.20), knowledge (0.16), understanding (0.11)
office hour	office (0.38), hour (0.38), time (0.09), comment (0.08), meet (0.08)
personality	enthusiastic (0.30), passionate (0.22), person (0.19), care (0.18), funny (0.12)
recitation	recitation (0.26), time (0.14), project (0.20), group (0.20), organized (0.20)
skills	knowledgeable (0.40), experience (0.26), information (0.14), quality (0.10), deep (0.10)
teaching assistant (TA)	TA (0.41), interactive (0.15), supportive (0.15), constructive (0.15), feedback (0.15)

(b) Highest ranked words list for each aspect of *Comments on overall assessment of this instructor*.

Table 4: Highest ranked words and normalized weight for each aspect.

Terminology	Description	Example
general feeling	General high-level comments or overall feelings about the course	The course is really interesting for me as a CS major and I learned a lot form it.
instructor	Any comments towards the instructor	Professor [NAME] is a joy, and is incredibly understanding, passionate, and enjoyable to simply listen to in class!
teaching assistant (TA)	Any comments related to TA	Resources are always available, the instructors and TAs were easily accessible and always friendly, the material was challenging, and examples were always fun and engaging.
lab	Any comments related to lab	I thought this course was very engaging and I liked that it was very hands on, like a lab should be.
recitation	Any comments related to recitation	This recitation was a bit odd.
course design	Any comments on the organization and structure of the course	I thought this course was excellently structured and formatted.
assignment	Any comments related to homework/assignments	The assignment is really, really well designed that it builds upon each other from assignment 2 through assignment 9 and it helped me exercise various topics/concepts that I learned from class.
exam	Any comments related to exam/test	This class is extremely hard and the second test is expected to be failed by most students, which is ridiculous.
content	Any comments related to course materials or specific contents of the course	The material was very useful for our course although the professors could have made a better connection with the techniques learned in the lab.
participation	Talk about the participation / attendance / engagement / discussion	Needs more class participation and discussion.
grade	Comments on the grading of the course	Harsh grading on lab reports.
group work	Any comments related to group work	All of the recitations consisted of group work towards a final project, though the early recitations seemed largely irrelevant to the project.
resources	Resources provided by course such as readings, textbooks, peer tutors etc.	I did however get all the help I needed from the peer mentors.
lessons learned	Learning outcomes or skills acquired from the course	The professor is really good, I learned a lot of interesting and classic dramas this semester.

Table 5: Aspect Annotation Terminology for *Comments on overall assessment of this course*.

Terminology	Description	Example
general feeling	General high-level comments or overall feelings about the instructor	Awesome Professor.
teaching assistant (TA)	Any comments related to TA	One of the best TAs I have had so far at UNC
recitation	Any comments related to recitation	Recitation felt like a waste of time.
office hour	Any comments related to office hour	She was great during her office hours and was always concerned that we understood the material.
personality	Describe personality of the instructor	She cares a lot about the subject material and her students.
skills	Describe the skill sets or experiences of the instructor	The instructor had a deep understanding of the course material and provided many real world examples built from her own experience and previous work.
grade	Comments on the grading style	The instructor was clear at explaining information and fairly evaluating all assignments.
delivery	How the instructor delivers the information and explains concepts	I really enjoyed her teaching style, she helped us through tough topics by breaking them down into more digestible chunks and was really positive overall, which helped for class moral.
course design	Comments on the organization and structure of the course	I didn't really get to know the TA because we didn't have a lot of recitations.
lessons learned	Learning outcomes or skills acquired from the course	I now have a greater understanding of the German language and of Swiss;German literature and culture.

Table 6: Aspect Annotation Terminology for *Comments on overall assessment of this instructor*.