IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 July 2023, accepted 10 August 2023, date of publication 15 August 2023, date of current version 24 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305260

APPLIED RESEARCH

Summarizing Students' Free Responses for an Introductory Algebra-Based Physics Course Survey Using Cluster and Sentiment Analysis

HONGZIP KIM[1] **AND GETING QIN**[2]
[1]Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada
[2]Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada

Corresponding author: Geting Qin (janice.qin@mail.utoronto.ca)

This work was supported by the Research Opportunity Program (ROP) with the University of Toronto, St. George campus under the supervision of Dr. Carolyn Sealfon.

Hongzip Kim and Geting Qin are co-first authors.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was approved by the Research Ethics Board (REB) of the University of Toronto.

ABSTRACT In Physics Higher Education (PHE), Student Evaluation of Teaching (SET) surveys are widely used to collect students' feedback on courses and instructions. In our research, we propose a more efficient way to summarize students' free responses from the Student Assessment of their Learning Gains (SALG) survey, a form of the SET survey, of an algebra-based introductory physics course at a large Canadian research university. Specifically, we use cluster and sentiment analysis methods such as K-means and Valence Aware Dictionary for sEntiment Reasoning (VADER) to summarize students' free responses. For cluster analysis, we extract popular keywords and summaries of responses in different clusters that reflect students' dominant opinions toward each aspect of the course. Notably, we obtain an average silhouette coefficient of 0.480. In addition, we analyze sentiments in students' free responses that are determined through applying VADER. Intriguingly, we see that VADER (micro F1 = 0.57, macro F1 = 0.55) can better classify responses with positive (F1 = 0.62) and neutral sentiment (F1 = 0.59). However, evident disagreements arise with negative sentiment responses (F1 = 0.42). In addition, our research suggests that some Likert-scale summaries deviate from the sentiment of free response summaries due to the limitations of Likert-scale responses. By creating various visualizations, we discover that Natural Language Processing (NLP) methods, such as cluster and sentiment analysis, effectively summarize students' free responses, with several limitations.

INDEX TERMS Cluster analysis, education, free responses, sentiment analysis, summaries, survey.

I. INTRODUCTION

For several decades, Student Evaluation of Teaching (SET) surveys have been widely implemented in Physics Higher Education (PHE) [1]. They generally consist of Likert-scale and free response questions that allow students to express their opinions on different aspects of the course [2]. For many years, students' free responses are usually understood and analyzed through manual reading [3]. However, there are often cases where little attention is paid to free response SET survey comments as analyzing vast amounts of free response data takes time. Therefore, there is often a lack of analysis or meaningful reports of free response comments, despite their importance [1].

In this research paper, we propose an effective way of summarizing students' free responses from the Student Assessment of their Learning Gains (SALG) survey [4] from an introductory algebra-based physics course at a large Canadian research university. By applying Natural Language Processing (NLP), we summarize free responses to provide a convenient way for instructors to quickly and effectively understand and interpret the survey results. Specifically,

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

we apply cluster and sentiment analysis to determine how effective these analyses are in summarizing students' free responses, enabling instructors to obtain a more holistic picture of students' thoughts toward the course and further improve the course design and implementation. Prior study shows that a more inclusive learning environment can take place when teaching shifts from a discipline-centered perspective to a student-centered perspective [5]. Student feedback on how the course is conducted and how much knowledge they have obtained would thus be helpful for instructors to improve the quality of instruction and learning environment for students. Additionally, qualitative feedback is critical because it often provides specific information and suggestions to improve teaching [1]. In other words, student evaluations are one of the necessary components in providing a more student-centered approach in PHE.

However, there are also some limitations to SETs. For instance, research suggests that students' learning is weakly correlated to SETs because students only assess the teaching quality of instructors instead of reflecting on the perception of their learning gains. The same study also suggests that students do not learn more from professors with higher SET ratings [2]. These findings by Uttl et al. concerning the weakness in correlations of SETs are significant to their limitations because they highlight that SETs are primarily measures of student satisfaction [2]. Specifically, students' responses can vary depending on factors that have nothing to do with the instructor's teaching effectiveness, also known as teaching effectiveness irrelevant factors (TEIFs). These can include student interest, the field of study, student motivation, instructor accent, and many more [2]. Studies suggest that instructors might focus more on increasing their rating from SETs rather than improving their teaching. Moreover, results from SETs are often ignored by faculty members due to their concern about the validity of SETs. Thus, collecting students' responses from SETs might not always facilitate improvement in teaching [6], [7].

Despite the limitations of SETs, they still provide a platform for students to express their opinions and valuable insights on the course for instructors, which allows for mutual benefit when instructors analyze students' free responses. However, to address the limitations of standard SET surveys, the SALG survey avoids assessments of the instructor and their performance as these factors are detached from students' perceptions of their learning gains [4]. As a result, we explore various ways of analyzing students' free responses from the SALG survey.

A. RESEARCH QUESTION

Prior research suggests numerous ways of summarizing and interpreting students' free responses, more of which would be elaborated in the Related Works section. Notably, Feng et al. use K means to cluster students' data and analyze their academic performance [8]. Additionally, Almatrafi and Johri use VADER as a tool for sentiment analysis in analyzing learners' responses to Massive Open Online Courses (MOOCs) [9]. The authors test the accuracy of VADER using the kappa coefficient to obtain a 0.41 kappa score, which is a good overall score. These approaches have supporting evidence showing that they are beneficial in extracting students' free responses.

As previously introduced, SET surveys have been widely used in PHE due to the importance of receiving student feedback. However, little attention is paid to free response SET survey comments, as analyzing vast amounts of student responses is time-consuming, resulting in a lack of analysis [1]. As a result, we seek to build on previous work and construct a straightforward approach to summarizing responses that do not require a deep understanding of machine learning, making the analysis of free responses more accessible and feasible for instructors from different fields around the globe.

For this reason, we investigate how we can effectively summarize students' free responses to SETs. As we aim to find a beneficial approach to analyze students' free responses for instructors through the utilization of cluster and sentiment analysis, we use cluster analysis to extract meaningful keywords from students' responses and give instructors a way to draw the most valuable feedback using the result of cluster analysis. We also extract summaries from each cluster as a way for instructors to have specific summaries related to a given course area.

Additionally, we discover how students' attitudes and sentiments from their free responses agree with their Likert-scale responses through sentiment analysis. By examining the difference between the free-response answers in conjunction with students' answers to Likert-scale questions, we also discover whether the median sentiment of free responses aligns with the median of Likert-scale responses. Comparing Likert-scale responses to free responses enables us to see whether free responses or Likert-scales can have more information about a student's perceived learning gains in the course. More details on how we achieve this are discussed in the Data & Methods section. Overall, reading various peer-reviewed articles, pondering our reasoning for such methods, and identifying the problem leads us to our main research question:

To what extent is cluster and sentiment analysis effective for summarizing and analyzing the Student Assessment of their Learning Gains (SALG) survey in algebra-based physics courses at a large Canadian research university?

A subset of our research question is:

How effective are NLP summaries of free responses compared to the statistics from Likert-scale responses?

The next consecutive sections of our paper are as follows: Background, Related Works, Data & Methods of our research, followed by the Analysis, Discussion of our results, Limitations, and our Conclusion.

II. BACKGROUND

Cluster analysis is an unsupervised technique for analyzing data which identifies groups or clusters of data points

that share similar characteristics. These clusters are formed from the natural patterns or structures within the dataset. In this discussion, we use K-means, which is a clustering algorithm that partitions M data points in N-dimensional space into K clusters based on their similarities. K-means operates repeatedly by assigning each data point to the nearest cluster centroid and updating the centroid of each cluster based on the mean of the data points assigned to it [10]. K-means aims to minimize the within-cluster sum of squares, which measures the variability of the data within each cluster. Consider the following objective function for K-means [11]:

$$W = \sum_{j=1}^{K} \sum_{i=1}^{M} ||x_i^{(j)} - c_j||^2 \tag{1}$$

where K is the number of clusters, M is the number of data points, and c is the centroid.

By minimizing the objective function (1), the algorithm ensures that the data points within a cluster have more similarities compared to other clusters. In fact, K-means is a widely used technique that finds applications in data reduction, data visualization and grouping [12].

Moving on, sentiment analysis is an opinion-mining technique that analyzes people's opinions, emotions, and attitudes of their responses, based on a computational factor of the subjectivity of the text response [13]. In our paper, we use the Valence Aware Dictionary for sEntiment Reasoning (VADER), a lexicon-based sentiment analyzer which provides a sentiment score based on the attitude and emotions carried out in the words. VADER uses qualitative and quantitative techniques to generate a sentiment lexicon, a standard set of words or phrases that indicate positive, negative, or neutral sentiment [13]. The lexicon from VADER is specifically designed to be effective in microblog-like contexts. Although VADER is specifically designed for social media texts, previous research has shown the effectiveness of VADER in analyzing the sentiment of student responses [13]. One of the main outputs from VADER is the sentiment score of a given text. The sentiment score can range anywhere from -1 to 1, where -1 signifies the most negative, 0 signifies the most neutral, and 1 signifies the most positive. The sentiment score is determined based on the compound score, (denoted as "x"), which is calculated by a normalized metric which incorporates the sum of all the valence score for each word in the VADER lexicon to be a score between -1 and 1, where:

- pos sentiment: $x \geq 0.05$
- neg sentiment: $x \leq -0.05$
- neu sentiment: $-0.05 < x < 0.05$

This classification of positive, neutral, and negative sentiment is based on the original VADER paper, where the authors discuss the classification of sentiment based on the compound score [13].

Both types of prominent NLP methods have been used in prior research associated with summarizing student feedback, but there is a lack of focus on research surveying students in Physics Higher Education (PHE). Hence, we will apply these two prominent NLP methods and determine the effectiveness of these methods with students' responses in an algebra-based physics course setting.

III. RELATED WORKS

Student surveys are often a key component for receiving students' feedback towards a course in physics education. Notably, the idea of improving and examining students' learning has been widely done since the 1990s [14]. For instance, Hake surveys the pre-test and post-test data of the Halloun–Hestenes Mechanics Diagnostic test, also known as the Force Concept Inventory, in classrooms that implement the interactive engagement (IE) methods, and traditional classrooms which did not implement IE methods. The results indicate that students in the classroom with IE-integrated methods perform better on average than students in a traditional classroom, demonstrating the feasibility of a survey for analyzing students' learning [14]. In fact, in 2004, a study by S. Freeman et al. finds that students in active learning classrooms have an average examination score that is 6% higher than students in traditional classrooms [15]. A study from Ornke et al. shows that students' success in learning physics is impacted more by students-related factors such as lack of motivation and interest [16]. In more recent years, the work of Bray and Williams demonstrates that first-year students are affected by fear and pressure more than study skills when being asked about their perceptions of physics [17].

Additionally, analyzing students' free responses has been frequently done with various techniques. For instance, semi-structured and face-to-face interviews are conducted to analyze students' verbal feedback through manual coding [18]. Research from analyzing students' free responses from the SET survey suggests that SET ratings and students' learning are weakly correlated due to students' inability to accurately assess an instructor's teaching effectiveness [2]. As time progressed, researchers began to notice the power of text mining and NLP, particularly in analyzing free text data. Since then, there has been an increase in studies that are centered around analyzing students' surveys with respect to those techniques. For instance, Brauwers and Frasincar perform a comprehensive survey reviewing different sentiment analysis algorithms for Aspect-Based Sentiment Classification (ABSC), in which the authors offer a categorization for ABSC models, including knowledge-based, machine learning, and hybrid models [19]. Additionally, in the work of Ferreira-Mello et al., the authors summarize and review the application of text-mining techniques, and highlight the adoption of NLP in educational platforms [20]. In 2020, Hujala et al. use Latent Dirichlet Allocation (LDA), a topic modelling technique, to extract and provide validated summaries of student evaluations of teaching and suggests that topics extracted from free responses can provide unique information not covered by Likert-scale questions [21]. Lastly, Schouten et al. propose a Java framework, "Heracles," for constructing and evaluating text mining algorithms [22]. The framework consists of the

NLP aspect for processing text, the Machine Learning aspect for text mining and an evaluation process for the designed text-mining algorithms. As a result, there has been an increase in studies focusing on such ideas.

Moving on, NLP algorithms such as speech tagging, cluster and sentiment analysis are used to visualize the free response sections of Student Evaluations of Teaching (SETs). In particular, the application of cluster analysis in education has slowly grown in popularity in recent years. In 2021, Delgado et al. proposed an unsupervised clustering technique based on neural networks to analyze 1,709,189 students' data across the span of four years, which shows that peer interactions are highly correlated with students' performance [23]. In the work of Cummingham-Nelson et al., pictorial visualizations of students' sentiment in the responses to survey questions are shown to be helpful for instructors to have an overall idea of students' perceptions toward the course. However, limitations arise as educators are concerned about the accuracy of sentiment analysis and the validity of SET surveys [1]. In 2017, Aung and Myo introduced a lexicon-based approach to analyze students' sentiments using an English sentiment word database as the lexical source [24]. In the same year, Alblawi and Alhamed [25] provided a comprehensive discussion of the use of NLP in higher education in their work. With their best model for sentiment analysis, they achieved a coefficient of determination (R-squared) of 0.89, which shows high accuracy. In addition, prior research suggests that Palaute is feasible for practitioners and provides accurate feedback [3]. Palaute incorporates cluster and emotion analysis to help instructors extract information from students' free responses. Lastly, Altrabsheh et al. proposed a real-time sentiment analysis method for analyzing student feedback using Support Vector Machine (SVM) and achieved an accuracy of 95% when ''neutral sentiment'' is not included as one of the possible sentiments in the classification process, since the class size for neutral was significantly lower compared to other classes [26].

As a result, all of these works contribute to our ideas for this research as we see little attention being paid to extracting and summarizing students' free responses from other surveys besides the traditional SET surveys. In this context, the traditional SET surveys only consist of questions centered around the quality of instructions. Hence, by building on previous research, we decide to take another step forward using NLP methods such as cluster and sentiment analysis to summarize and analyze students' responses from the SALG surveys.

IV. DATA & METHODS

A. DATA

To begin with, this research is conducted in full compliance with the Research Ethics Board (REB) of the University of Toronto, where data is anonymous, reported only in aggregate and password-protected. Consent to accessing the data was given once the training, TCPS 2: CORE-2022 (Course on Research Ethics), was completed. As a result, to ensure the privacy of the students and to follow the regulation of the REB, student data will not be available to the public, but the code for this study will be published for reproducibility purposes. In addition, the SALG survey is also publicly available on their official website.

In terms of data collection, we use the SALG survey data from the winter semesters of an introductory physics course at the University of Toronto, St. George campus, from 2019 to 2021. The SALG survey data contains both Likert-scale and free response questions. Notably, it is an online survey which students fill in at the end of the course using a computer, and students' responses are received on the SALG website afterwards. The instructors receive all of the student feedback via an Excel file. It aims for students to reflect on their perception of learning gains in various aspects of their studies. Different from traditional SET surveys, this survey does not contain questions that assess instructors' teaching ability that is irrelevant to students' perception of their learning gains [4]. Therefore, with survey questions designed explicitly for students to reflect on what they have learned in class, the SALG survey helps to assess students' perception of their learning directly, in which feedback and suggestions are collected for instructors to improve their course facilitation. In particular, students are allowed to express their thoughts freely, as there are no limitations on the number of words that students can answer for each free response question. The survey also helps assess students' perception of their learning outcomes, as it contains several sections asking students to reflect on these areas. Referring to Table 1, with a comprehensive list of the SALG survey sections, as well as abbreviations that will be used onward, sections like ''Your understanding of class content,'' ''Increases in your skills,'' and ''Integration of your learning'', all place emphasis on assessing the learning outcome of students. Therefore, we can see the students' perception of their learning outcomes when analyzing the survey results.

Approximately 1,000 students answered the SALG survey from 2019 to 2021, and there are around 15,000 free responses, which can be shown in the following calculation:

f = the total number of responses
x = number of students participating in three years
y = number of free response questions in the survey

$$f \approx xy \approx 1,000 \times 15 = 15,000$$

However, it is important to note that not all students responded to every question, so 15,000 is only an approximation. Since we only use valid responses, we only analyze 12,513 student responses across all three years of the SALG data. Responses that are ''NA'', ''n/a'', ''.'', blank spaces, and so on count as invalid responses. Therefore, after filtering out these invalid responses, the data set is reduced to 12,513 responses. Filtering out these invalid responses does not affect our research, as they do not contain any reflection and feedback.

In terms of the SALG survey, each section contains at least 1 free response question and no more than

15 Likert-scale questions. The responses for each section of the Likert-scale questions are from 1-5, representing "no gains/ help" to "great gains/ help." Students can also select "not applicable" in the survey, which is recorded as "9" in the data. These responses are not counted in calculating the mean and median of the Likert-scale questions included in the SALG survey data. The analysis of the SALG survey data is discussed in greater detail in the Analysis section. This data is highly relevant to our research topic and focus, as we demonstrate a feasible approach to summarize learners' free responses on SETs and compare the sentiment of free responses to answers for Likert-scale questions.

TABLE 1. Different sections of the SALG survey and their corresponding abbreviation. We used these abbreviations in Fig. 4, Fig. 6, and Table 7.

Section	Abbreviation
Your understanding of class content	understanding
Increases in your skills	skill
Class impact on your attitudes	attitude
Integration of your learning	integrate
The Class Overall	overall
Class Activities	activities
Assignments, graded activities and tests	assignments
Class Resources	resources
The information you were given	info
Support for you as an individual learner	support
Improving the course	improvement

B. METHODS

1) PRE-PROCESSING SALG SURVEY DATA FOR CLUSTER ANALYSIS

First, we extract all of the free responses for each year of the SALG survey data, which only contains students' complete responses (omitting responses such as "N/A", "/", "n/a", "None", "none", ".", "−", "N", "Nope", "nope", "No", and blank spaces). We clean the text by making all the words inside the data frame lowercase, tokenizing the text and removing English stopwords using the Natural Language Toolkit (NLTK) [27] in the data frame. Afterwards, we convert each word into a vector by applying the term frequency-inverse document frequency (TF-IDF) [28] algorithm and setting max_df = 0.95. By doing so, the result from TF-IDF does not include words that appear in 95% of the document, eliminating unnecessary frequent words that might not be counted as stopwords.

2) PRE-PROCESSING SALG SURVEY DATA FOR SENTIMENT ANALYSIS

To prepare for sentiment analysis, we extract the free responses and the Likert-scale questions and copy them over to a new Excel file. Then, we delete the first row as it corresponds to the headings of each question (1.1, 1.2, 1.3, 1.4, . . .). We save and name the file as winter20xx_<"section name">, where the section names can be seen in Table 1. We repeat this for each section (10 total) for the three years' worth of data (10 × 3 repetitions).

Afterwards, we read the Excel file and loop through each row, as well as the cells within the row. If the length of the cell is "1", then the Likert-scale response casts from a "string" to an "int." However, when the response consists of dashes, periods, slashes, and other non-numerical responses, they are filtered out so there are no errors in casting. Otherwise, the cell's content is stored in a list. Finally, we filter out the responses such as N/A, and None. In addition, if the Likert-scale responses contain a "9", where 9 means N/A, then those students' entire (both Likert and free) responses are omitted. In other words, we only keep Likert and free responses from students who fully answered both survey segments. As a result, our results are derived from only a part of students' responses, hence making our sample not representative of a census of all student responses.

3) CLUSTER ANALYSIS

For cluster analysis, we first need to reduce the dimension of our data as the data is highly sparse and in high dimensions. We apply two types of dimension reduction algorithms to achieve the optimal visualization result of data in high-dimensional space. To begin with, we use principal component analysis (PCA) [29] to reduce the dimensions of the clustering result data to 10 by converting the result of TF-IDF (which is a highly sparse matrix) into an array and setting the n_components input parameter of PCA to 10. Then, we use t-distributed stochastic neighbouring embedding (t-SNE) [30] to further reduce the dimension of data points to 2 by setting the input parameters of t-SNE as the following: n_components=2, verbose=1, perplexity=30, n_iter=5000, learning_rate=100. The reason we combine PCA and t-SNE is that using only one will not achieve optimal results. If we only use PCA, then the graph does not showcase distinct clusters and the data points are cluttered. On the other hand, if we do not use PCA and only use t-SNE, then we get a pile of data points clumped in the center of the graph, as t-SNE does not perform as well when applying it to high dimensional data (dimensions greater than 50) [31]. Therefore, we select the components of PCA to be a dimensional space of 10 after looking at the t-SNE result with different components of PCA.

Once the dimensions of the data are reduced, we apply K-means to cluster the vectorized students' responses. As mentioned before, LDA is also frequently used for analyzing large amounts of text responses, but it only outputs topics in the text since it is a topic modelling technique. Therefore, the result from LDA could potentially convey less information than cluster analysis. We choose K-means as our clustering technique as it is a well-known unsupervised clustering technique which has been used for clustering survey results in the past.

Due to its unsupervised nature, it is easy for instructors with foundational Python knowledge to implement by themselves.

Initially, we attempted to use the elbow method for selecting the number of clusters in the early stages of this experiment but found no distinct elbow. Therefore, we manually select the number of clusters (K) based on the visualization results of a range of 5 - 15 clusters as doing this allows for a more insightful analysis. Specifically, we selected the number of clusters based on the plot for the data after dimension reduction. For the 2019 SALG survey data, we select $K = 12$; for the 2020 data, we select $K = 11$; and for the 2021 data, we select $K = 12$. Since K-means is a stochastic clustering algorithm, it can produce different results from different runs even if the input data stays the same. To ensure the result from K-means clustering is reproducible, we set the random_state of the algorithm to be 50 [10]. To check the validity of the cluster results, we use the silhouette coefficient [32], which is a metric between -1 to 1 that represents the goodness of cluster results, to determine the quality of our clusters. As a final step, we store and plot the clustering result. To determine the topic of each cluster and label the cluster accordingly, we extract the words with the top 30 highest TF-IDF values in each cluster and manually determine the cluster label.

For summary extraction, we trace back the sentences contained in each cluster and determine the frequency of each word in the sentences. Here, we count the number of words appearing in the given text, which is different from TF-IDF, because this approach is relatively straightforward for instructors with no background in NLP but would like a quick and easy way to generate summaries of students' responses. The output of TF-IDF is a highly sparse matrix for our data (with around 4000 rows and columns) due to the number of students' responses. As a result, TF-IDF is harder to work with compared to a Python dictionary with words and their corresponding frequency.

Therefore, we extract a summary for each cluster by creating a dictionary of words (omitting stopwords) in the cluster and their corresponding frequency, which is achieved by counting how many times they appear in the given text. Then, for each response in the given text, we add up the total frequency of words and store the response and its total frequency. We compute the average frequency for all responses by the following:

- Let sumValues be the value of the summation of the frequency of each response.
- Then, average frequency = sumValues / number of responses in the given text.

We take the average because if we add up all the frequencies of the words in the response, a longer response will tend to have a higher frequency. However, it does not necessarily represent that a particular response is a general representation of the other responses. Hence, to account for the different lengths of the responses, we decide to take the average word frequency in the response instead of using the sum of word frequency.

To create the summary, we extract responses with a frequency above a $1.3\times$ average frequency threshold. We use the average frequency of responses because we can use this as an indicator of how frequent the response is compared to other responses. The number "1.3" is manually selected based on the summary length we want. When it comes to practice, it depends on how much an instructor wants the responses in the summary to be more frequent than the average frequency. The higher the number means there are fewer responses in the summary. In the end, we manually edit the minor grammar mistakes in the summaries to strengthen the readability and indicate any modification with a square bracket.

4) STATISTICAL ANALYSIS

To compare the Likert-scale responses with the free responses using sentiment analysis, we first perform statistical analysis. We first convert the Likert-scale responses, ranging from 1 to 5, to a scale of -1 sto 1, as the sentiment scores range from -1 to 1. The conversion serves to reflect the sentiments portrayed in students' responses on their perception of their learning gains when compared to Likert-scale responses. Please see Table 2 for the conversion, as well as the interpretation of each Likert-scale response.

TABLE 2. Conversion of Likert-scale responses to the sentiment score scale of -1 to 1.

Likert-scale	Score Conversion	Meaning
1	-1.0	no gains
2	-0.5	a little gain
3	0.0	moderate gain
4	0.5	good gain
5	1.0	great gain

Then, we calculate the median of Likert-scale responses in each survey section. This is achieved by appending all the Likert-scale responses into one list and using the statistics.median() function in Python to sort the list in ascending order, and extract the median. Next, we extract all the preprocessed responses, and group the Likert-scale responses by the students. In the end, we visualize the differences between the Likert-scale questions and free response sentiments.

5) SENTIMENT ANALYSIS

For the implementation process of VADER, we first import VADER using the nltk.sentiment.vader package [27] and import the SentimentIntensityAnalyzer. By using the SentimentIntensityAnalyzer(), and inputting the cleaned student responses, we analyze the polarity score of the sentiment for each response, which consists of a dictionary of the percentage of positive (pos), negative (neg), and neutral (neu) sentiment, as well as the compound score. We extract the compound score from the dictionary and append it to our list.

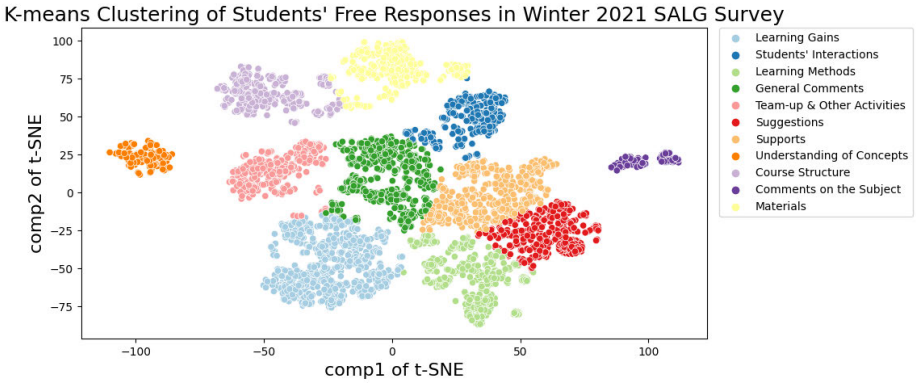


FIGURE 1. K-means clustering of physics students' free responses in the winter 2021 SALG survey with K = 12. Each colour in the figure represents one cluster, so responses in the same cluster have the same colour. Please refer to Table 4 for the corresponding keywords in each cluster. The "comp" in the axes labels is the abbreviation for the word "component", where comp1 and comp2 represent the axes for the two-dimensional space that the data are in after dimension reduction.

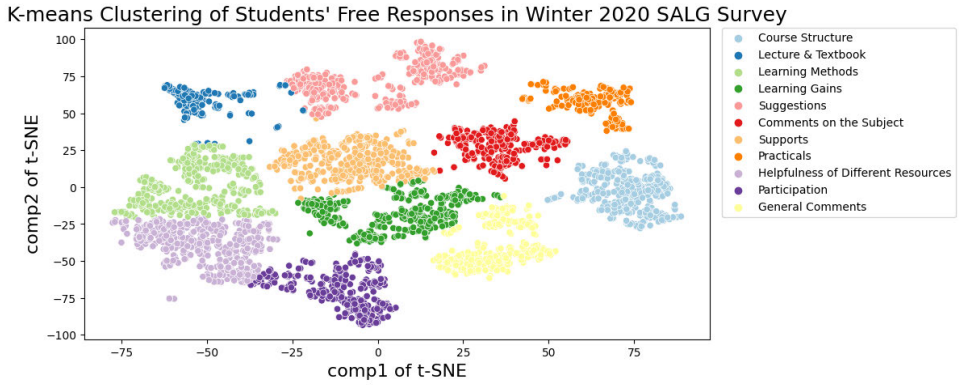


FIGURE 2. K-means clustering of physics students' free responses in the winter 2020 SALG survey with K = 11. Each colour in the figure represents one cluster, which means responses in the same cluster have the same colour. Please refer to Table 10 in the Appendix for the corresponding keywords in each cluster.

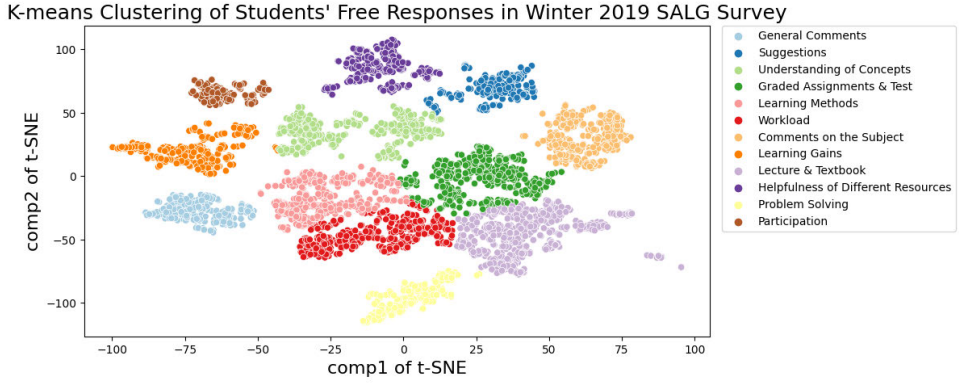


FIGURE 3. K-means clustering of physics students' free responses in the winter 2019 SALG survey with K = 12. Each colour in the figure represents one cluster, so responses in the same cluster have the same colour. Please refer to Table 11 in the Appendix for the corresponding keywords in each cluster.

We also make another function to return the percentage of positive, negative, and neutral sentiment responses:

- (pos/neg/neu sentiment count) / (total number of responses)

TABLE 3. Example sentences of responses for each keyword from Cluster 10: "Online Class" of the 2021 SALG survey data. We extract these sentences by creating a function that returns the sentences that contain the corresponding keyword in the desired cluster. We randomly select two sentences as examples for each keyword.

Keywords	Examples
synchronously	1) due to time difference issues attending class synchronously was difficult. The team up activities could have had a longer due date (eg until the end of the week) 2) I used to only watch lectures asynchronously (because class times conflicted with other classes or work) but after I started watching classes synchronously, I found myself understanding subjects better and participating in active learning more.
Zoom	1) Last year, in person discussions with class mates were easier than the zoom breakout rooms this year. 2) The only help I received was from group chats outside of the class zoom chats because my questions always went ignored. The in class problems were never helpful because the instructor had disorganized ways of solving, she just writes everywhere all over the screen not making it easy for anyone to follow along
discussions	1) I will always take an active part in class discussions. 2) I don't think I would participate very actively since I'm expecting it to be a very big class, and I don't feel comfortable speaking in front of a big crowd. However, I may participate more in the small group discussions since there are more means for me to share my ideas instead of just talking.
participated	1) I will always participated in class discussions with others. 2) I rarely participated in the class discussions because people did not talk frequently
atmosphere	1) I would sometimes participate in class. The class had a very welcoming atmosphere which encouraged participation. 2) rarely, the atmosphere in the classroom during class activities was inviting seeing the homework was a group activity.

Lastly, we separate the cleaned student responses into free responses longer than or equal to one sentence. Using the results we obtain from statistical and sentiment analysis, we visualize the results with various graphs. Particularly, we create:

1) a side-by-side box plot for each survey section (see Table 1 for the sections in the SALG survey. We omit the section named ''improvements,'' as it does not contain Likert-scale questions in the survey),
2) another side-by-side box plot for free responses that are longer than one sentence or equal to one sentence, and compare their Likert-scale and sentiment score distribution, and
3) a stacked bar plot which shows the percentage of pos/neg/neu sentiments in each survey section.

Finally, to test the accuracy of VADER, we extract and manually label 800 free responses, so that we can get the F1 score for VADER. We randomly extract 100 to 300 sample responses at a time, until we reached a total of 800 distinct responses, and then manually label the sentiment of each response. In particular, both authors were involved in the labelling process as annotators with an Inter-Annotator Agreement (IAA) kappa score of 0.74, which indicates a substantial agreement between the annotators. Using the data, we calculate the micro and macro F1 score by calculating the VADER sentiment score of the free responses and determining whether it is positive/negative/neutral based on the compound score classification. The F1 score is a harmonic mean of precision and recall, where precision is the ratio of true positive with the sum of true positive and false positive, and recall is the ratio of true positive with the sum of true

TABLE 4. The keywords in the clusters and their corresponding topic for the 2021 SALG survey. The topic for each cluster for this year might not indicate that there are similar contents of responses from year to year with the same topic. These topics are labelled based on the keywords of each cluster for the 2021 SALG survey data. They are not cross-referenced to match the cluster labels from different years.

Cluster	Topic	Number of Student Responses	Keywords
0	Course Structure	333	students, team, time, center, resources
1	Learning Gains	722	interest, ability, gained, problems, situations
2	Suggestions	416	watched, improve, confusing, focus, due
3	General Comments	512	subject, ideas, test, students, material
4	Team-up & Other Activities	350	mentioned, practical, approach, learned, study
5	Materials	364	perusall, resources, readings, content, lecture
6	Learning Methods	436	practice, tests, interrogating, format, image
7	Supports	501	piazza, information, group, ta, support
8	Students' Interactions	304	helped, practical, peers, concepts, useful
9	Understanding of Concepts	160	test, homework, questions, teamup, understanding
10	Online Class	114	synchronously, Zoom, discussions, participated, atmosphere
11	Comments on the Subject	316	information, honestly, material, phy, subject

TABLE 5. Summaries for three different clusters of students' responses across three years. These summaries consist of student responses that are in each cluster with a frequency higher than 1.3 times of the average frequency of the responses in the given cluster. Only the first half of the summary for "Team-up & Other Activities" is presented as this summary is longer than the other two.

Cluster and Year	Summary
Comments on the Subject (2019)	I don't think I will take more physics courses in the future, but the things I have learned in this course [are] very interesting. The teaching of the class relates physics concepts to real [phenomena], visually representing the knowledge and emphasizing my impression of the ideas. The instrumental approach of the class involved a lot of demonstrations to represent the application of the physics concepts being taught, and led me to approach the theories and formulas through logical thinking.
Learning Methods (2020)	I thought I could get everything only with lectures, and now I know that I should spend more time on reading textbooks in detail[.] I force myself to pay more attention before tests, and that's usually when I realized that [I] did not pay enough attention to details. I would discuss with the people around me, but the atmosphere wasn't something I actively engaged in, as it felt better to just discuss with people 1-2 seats away. Discussing and working with peers is also very helpful, for it motivates me to learn from others.
Team-up & Other Activities (2021)	Team-ups helped a lot to understand the quizzes because it had a similar goal[.] The way the TA would grade our practicals really helped [me] understand what was wanted from us[.] The support received from my classmates outside of class really helped [me] understand because we had access to different point[s] of views[.] Lots of real-world applications which is useful but I would have liked some more quantitative-based problems to solve in this class I learned a lot from the simulations than from the texts we had to read for homework.

TABLE 6. Silhouette coefficients for the clustering result of each year's SALG data. The average of the three years is 0.480, which indicates the clusters are reasonably separated.

Cluster Result	Silhouette Coefficient
Winter 2021	0.478
Winter 2020	0.474
Winter 2019	0.488

positive and false negative. A micro F1 score calculates the metrics globally by counting all the true positives, false positives and false negatives, whereas a macro F1 score calculates the F1 scores across all classes [33]. We choose to include both the micro and macro F1 scores as we want to calculate the proportion of correctly classified sentiments, which is essentially the equivalent of calculating the accuracy [34], and also to help us identify if the model is biased towards a certain class. Thus, given the nature of the dataset being imbalanced, incorporating both the micro F1 score and the macro F1 score provides a more all-inclusive measure of VADER's performance.

V. ANALYSIS

A. CLUSTER ANALYSIS

As shown in Fig. 1 & Table 4, the 2021 data showcases the implementation of "Team-Up!" (see the pink cluster), [35] which encourages student-to-student interactions during class time, differing from the previous two years. We see 350 students' responses in the "Team-up & Other Activities" cluster, which shows students' interest in discussing the activities in the course. Noticeably, a large proportion of responses (N = 722) are in the "Learning Gains" cluster, reflecting students' perceptions of their learning gains, which suggests that the survey could be well-designed as it achieves the purpose of assessing students' perceptions of their learning. The appearance of responses (N = 114) around "Online Class" may reflect how the outbreak of the COVID-19 pandemic has sparked students' discussion about the change of course delivery method. In 2021, the course was delivered synchronously, which means that students had a set time to attend lectures over Zoom. Keywords such as "synchronously" and "Zoom" in Table 4 all reflect students' discussion in the change of course delivery method.

Looking further into instances where the keywords take place in a student's free response, consider Table 3. We randomly select these examples for the cluster "Online Class" in the 2021 SALG survey data to showcase responses that contain each keyword. These examples suggest that the responses classified as "Online Class" are accurate for the 2021 survey data. We see that many responses revolve around how students feel about participating in class discussions over Zoom.

Referring to Fig. 2 & Table 10 in the Appendix, many of the responses revolve around students' comments on course

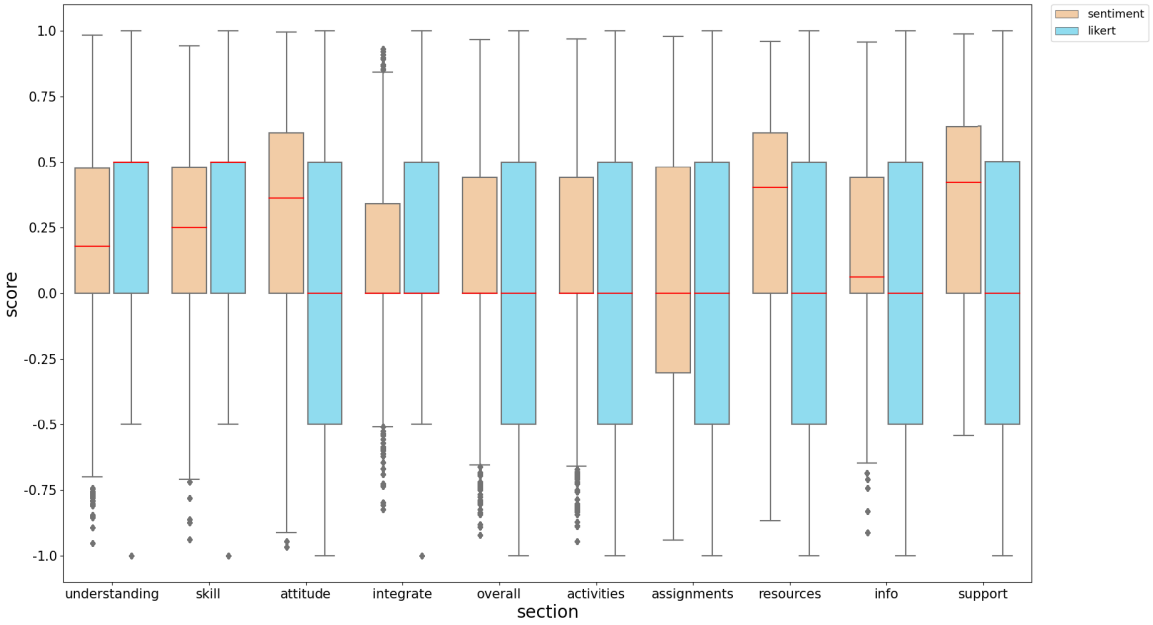


FIGURE 4. Visualization of side-by-side box plots comparing the VADER sentiment scores (in orange) and Likert-scale responses (in blue) for each SALG survey section (labelled as "section" on the x-axis). The medians of Likert-scale responses and sentiment scores are coloured in red. The left y-axis label is for the sentiment score. Please see Table 2 for the Likert-scale to sentiment score conversion chart, and the interpretation of each Likert-scale response.

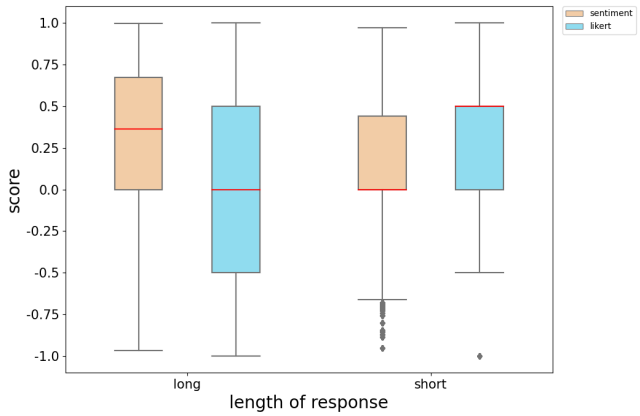


FIGURE 5. Box plot comparing the VADER sentiment and Likert-scale data for long (longer than one sentence) vs. short (less than or equal to one sentence) responses. The left y-axis label is for the sentiment score.

structure (N = 496) and around lectures and the textbook of the course (N = 538). Corresponding keywords such as "lecture, discussions", and "resources, topics" show the accuracy of the grouping between the labels and students' responses in these clusters.

As illustrated in Table 5, we showcase a few summaries we extract from clusters in the three years of SALG survey data. For instance, the summary for comments on the subject is displayed for 2019, where students share their thoughts on the subject itself and what they perceived to have learned in the course. In 2020, students talked about how the in-class activity and talking with peers help them when solving physics problems, which suggests the helpfulness of these learning methods. It also shows students' reflection on the improvement of their learning methods choices, as part of the summary explains how a student may recognize that solely going to the lecture might not be entirely effective for learning physics. For the summary of the cluster "Team-up & Other Activities" in 2021, we can see that the majority of students liked the implementation of activities such as "Team-Up!" when the course was delivered online in 2021. This summary reveals that students find these collaborative activities interesting and helpful, and these opinions in the summary suggest the instructor should continue implementing the "Team-Up!" activities.

Lastly, looking at Table 6, we can see that the silhouette coefficients for all three years of cluster results are above or equal to 0.474, with the average of the three years being 0.480. These silhouette coefficients indicate that the cluster results are reasonably separated, with room for improvement in terms of making the clusters more distinct and split from each other.

B. SENTIMENT ANALYSIS

From the box plot (Fig. 4), we see that the interquartile range (IQR) of the box plots varies for each survey section, suggesting there might be differences in information and sentiment of student responses conveyed in each survey section. Based on the box plot, we see that the medians of the data vary between the Likert-scale and sentiment scores. This is only an exception for "integrate," "overall," "activities," and

TABLE 7. Table comparing the sentiment and Likert-scale for each SALG survey section on various statistical data, including mean, standard deviation, minimum, Q1, median, Q3, and maximum.

Survey Section	Type	mean	std	min	25%	50%	75%	max
understanding	sentiment	0.214	0.371	-0.952	0	0.178	0.475	0.984
	likert	3.609	1.096	1	3	4	4	5
skill	sentiment	0.238	0.355	-0.938	0	0.25	0.478	0.942
	likert	3.604	1.126	1	3	4	4	5
attitude	sentiment	0.254	0.446	-0.967	0	0.361	0.612	0.994
	likert	2.986	1.329	1	2	3	4	5
integrate	sentiment	0.105	0.333	-0.823	0	0	0.348	0.93
	likert	3.396	1.151	1	3	3	4	5
overall	sentiment	0.153	0.376	-0.92	0	0	0.44	0.968
	likert	3.078	1.261	1	2	3	4	5
activities	sentiment	0.152	0.387	-0.946	0	0	0.44	0.97
	likert	3.219	1.357	1	2	3	4	5
assignments	sentiment	0.1	0.489	-0.94	-0.303	0	0.477	0.978
	likert	3.193	1.351	1	2	3	4	5
resources	sentiment	0.302	0.36	-0.866	0	0.42	0.599	0.961
	likert	3.26	1.291	1	2	3	4	5
info	sentiment	0.188	0.347	-0.911	0	0.076	0.44	0.957
	likert	3.054	1.213	1	2	3	4	5
support	sentiment	0.373	0.341	-0.542	0	0.422	0.637	0.988
	likert	3.219	1.402	1	2	3	4	5
improvement	sentiment	0.151	0.382	-0.952	0	0	0.44	0.983
	likert	N/A	N/A	N/A	N/A	N/A	N/A	N/A

TABLE 8. Table comparing the sentiment and Likert-scale for different lengths of responses on various statistical data, including mean, standard deviation, minimum, Q1, median, Q3, and maximum.

Response	Type	mean	std	min	25%	50%	75%	max
long	sentiment	0.25	0.479	-0.967	0	0.361	0.67	0.994
	likert	3.295	1.303	1	2	3	4	5
short	sentiment	0.173	0.342	-0.952	0	0	0.44	0.97
	likert	3.561	1.166	1	3	4	4	5

''assignments,'' where the medians line up at 0.00 sentiment (neutral), which converts to 3 for Likert-scale (some gains).

To understand the box plots better, we can look at Table 7, where we see that the minimum sentiment score is consistently around −1.00 to −0.80, with the one exception where the minimum sentiment score for ''support'' is −0.542, suggesting that students have a less negative perception of their support in classrooms. On the other hand, the maximum sentiment is always above 0.90, with the highest being 0.988 in ''support''. For the interquartile range (IQR), the 25th percentile (Q1) is almost always at 0 for sentiment scores (except for ''assignments'', where it is −0.303), and either 2 or 3 for Likert-scale responses. The median sentiment scores range from 0 to 0.422, but 0 is the most frequent median sentiment score across all the survey sections, which suggests that half of the students' responses carry a more positive sentiment and the latter half carry a more negative sentiment.

From the statistics, we see how much variance between the medians there is when comparing Likert and sentiment score distributions, which suggests that free responses may convey different information and attitudes than Likert-scale responses as students are free to write what they like instead of solely selecting numbers based on the questions in the Likert-scale section. As a result, comparing the distribution of Likert-scale responses and sentiment scores via a box plot potentially offers a feasible and quick approach for instructors to understand the different information that might be carried out in both free and Likert-scale responses. When examining the outliers that are in Fig. 4, all of the outliers happen to be responses with negative sentiment scores, with the sole exception of the sentiment score distribution of ''integrate.'' Upon further examination, we see that students are more extreme in their free responses when they are given the chance to elaborate on their thoughts. To investigate this prediction, we will move on to the next analysis: comparing the Likert-scale and sentiment score distribution of long vs. short responses.

Looking at the long (more than one sentence) vs. short (less than or equal to one sentence) response distribution (see Fig. 5 & Table 8) helps us understand the differences between Likert-scale responses and sentiment scores between varying response lengths. When we look at the Likert-scale and sentiment score distribution of ''long'' responses, we see that there is a more comprehensive range of sentiment score vs. Likert-scale distribution, where the IQR for both Likert-scale and sentiment scores for long responses are greater than short responses. It is further interesting to note that the maximum and minimum sentiment for longer responses is greater than those of one-sentence (short) responses, although they do not differ drastically. These results are consistent with our impression that students who elaborate on their free response tend to be more extreme in their emotion/sentiment as well as their Likert-scale responses.

Additionally, we see that from the stacked bar plot (Fig. 6), there is a higher percentage of positive responses compared to neutral and negative responses. We see that the

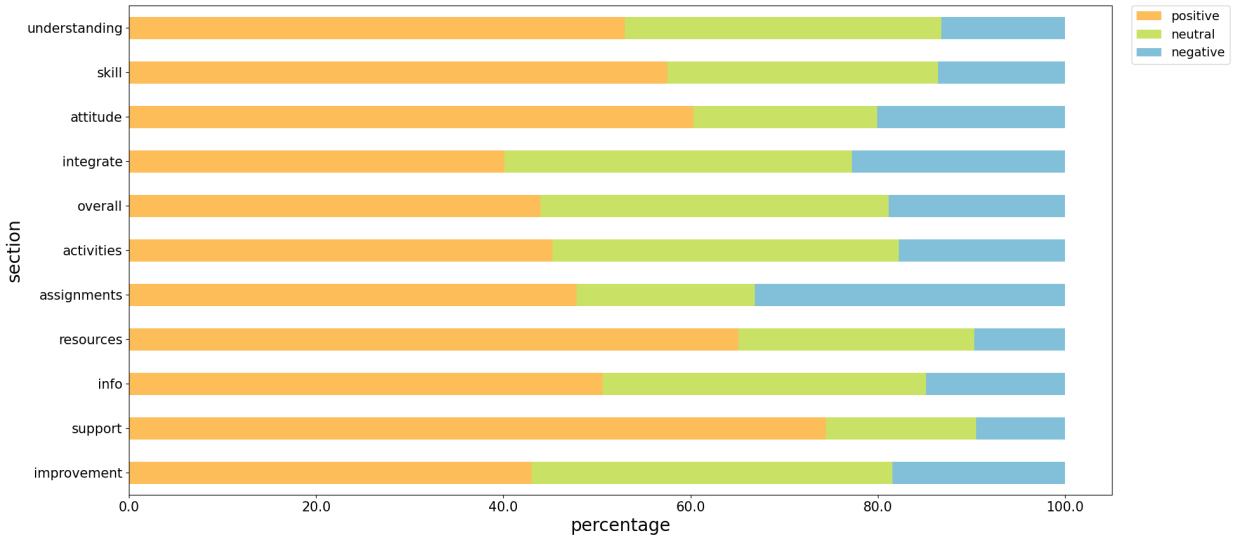


FIGURE 6. Stacked bar plot consisting of the percentage of "positive," (x >= 0.05) "neutral," (−0.05 < x < 0.05) and "negative" (x <= −0.05) responses for each SALG survey section, where x is the VADER compound sentiment score.

greatest percentage of positive responses are in "support," and "info." The percentage of negative responses is relatively low (<20%), with the exception where there are many "negative" responses in "assignments," followed by "integrate," and "attitude." Neutral tends to be the second most sentiment score, followed by "negative" for most sections. From this graph, we can observe the proportion of positive, neutral, and negative responses across all survey sections, and enable instructors to understand sections where students are the most or least satisfied.

To test the accuracy of VADER (Table 9), we calculate the F1 micro score, which is equivalent to the accuracy of classification (e.g. true positive, true neutral, and true negative), and the macro F1 score, which essentially computes the average F1 scores spanning all classes. Examining the table, we see that there are 397 responses which VADER classifies as "positive," and we only classify 264 responses as positive. Of those 264 responses, we only got 205 true positives, giving us an F1 score of 0.62 for positive sentiment responses. In addition, for negative responses, we see that there are 143 responses which VADER classifies as "negative," while we identify 187 responses as negative. Of those 187 responses, we only obtain 70 true negatives, giving us an F1 score of 0.42, which is the lowest of all three sentiments. Lastly, for neutral responses, VADER classifies 260 of those responses as "neutral," and we classify 349 responses as neutral. Of those 349 responses, only 181 responses are true neutrals, giving us an F1 score of 0.59. Combining all of them, we get an overall micro F1 score of 0.57 and a macro F1 score of 0.55. Our results suggest to us that VADER is better at classifying "positive" and "neutral" responses compared to "negative" responses. Examining the macro F1 score of 0.55, which is slightly less than the micro F1 score of 0.57, shows the possibility of some classes (e.g. positive, neutral)

TABLE 9. Confusion matrix of 800 sample student free responses and its corresponding micro F1 score results.

Actual	Prediction		
	Positive	Negative	Neutral
Positive	205	76	116
Negative	21	70	52
Neutral	38	41	181

Class	n (truth)	n (classified)	Accuracy	Precision	Recall	F1 Score
Positive	397	264	68.63%	0.78	0.52	0.62
Negative	143	187	76.25%	0.37	0.49	0.42
Neutral	260	349	69.13%	0.52	0.7	0.59

Overall accuracy:
F1 (micro): 0.57
F1 (macro): 0.55

performing better than other classes (e.g. negative), and the potential of dataset imbalance. This demonstrates consistency with previous research where VADER is also shown to perform better at classifying positive, and neutral sentiments compared to negative sentiments [9]. The results suggest that the positive and neutral classifications and their sentiment scores are more reliable to draw insights from, compared to responses classified as having a negative sentiment, with a compound score less than −0.05. As a result, instructors should be more careful in reviewing responses classified as having a negative sentiment, and be more cautious when interpreting it.

VI. DISCUSSION

A. CLUSTER ANALYSIS

From our results, we see how we can create multiple clusters using K-means for all three years of the SALG survey data. From these clusters, we manually extract keywords based on

the words that have the top 30 highest TF-IDF values. These keywords are then used to determine the ''topic name'' for each cluster. Furthermore, we can extract sentences containing these keywords from the keywords of a given topic, where the extracted sentences suggest that they are consistent with our labelling of the topic. Lastly, we extract summaries from each cluster in each year of the SALG data, which provide a holistic and simple representation of student responses in each cluster in a given year. The results from cluster analysis seem to be accurate with their silhouette coefficients all above 0.470, with an average silhouette coefficient of 0.480, showing the cluster results are acceptable for the purpose of this study. However, the number of clusters is manually chosen, which can be further improved. More details on how we plan to address this concern are discussed in the Conclusion section.

B. SENTIMENT ANALYSIS

From sentiment analysis, we see how we can compare the distribution of Likert and sentiment scores using VADER for each survey section in all three years combined. We then compare the distribution of Likert-scale and free responses that are longer than or equal to one sentence. We also look at the percentage of positive, negative, and neutral responses given a survey section, and test the accuracy between manual coding, with a and VADER via the F1 metric, using both the macro and micro averages, and a confusion matrix. From our results, we identify points of interest, such as the median and different percentiles, in the statistics associated with the Likert-scale vs. sentiment score box plots and find support for our conjecture that students' free responses vary more and perhaps convey more information than Likert-scale responses. The latter can be potentially explained by the fact that there is no word limit in the free response questions. To understand the reason for our prediction, we looked at comparing the distribution of ''long'' vs. ''short'' responses. These results suggest to us that students who elaborate more tend to be more extreme in their word choice/feelings and have a wider variety of Likert-scale responses compared to students who only write a sentence or less. Looking at the confusion matrix, we see how VADER is better at classifying positive, and neutral sentiments compared to negative sentiments. This result is consistent with findings from previous research [9].

As this process shows, by extracting summaries from various clusters and comparing the distribution of VADER sentiment scores and Likert-scale responses for each survey section, we are able to pilot a unique approach using NLP methods to create a feasible and time-efficient way for instructors to interpret survey free responses and attain a faster understanding of students' perception of their learning gains.

VII. LIMITATIONS

Although we have demonstrated how cluster and sentiment analysis can provide useful insights for instructors to analyze students' free responses, there are still potential limitations in this study. To start with, for cluster analysis, the selection of the number of clusters is done manually as the number of clusters suggested by the ''elbow'' method does not generate meaningful summaries. This, however, can be explained by the nature of our dataset being highly sparse and the high dimensional properties of the text data. Though a silhouette coefficient score above 0.470 is acceptable, it can still be improved further by exploring other cluster algorithms and investigating their results. In addition, in terms of sentiment analysis, the algorithm chosen for analyzing the students' sentiment, VADER, is trained on social media texts [13], which might differ from the students' writing tone and word choice. Though previous research has shown that VADER is fairly good at analyzing the sentiment of students' free responses, our research suggests that the results can still be further improved. One possible reason may be explained by the imbalance in the data set, which affects the outcome of the results. Based on the 800 random students' responses that are manually labelled, there are around 49.6% positive responses, 32.5% neutral responses and 17.8% negative responses. Hence, the imbalance in the dataset might be a factor that affects the accuracy of using VADER.

VIII. CONCLUSION

In classrooms, student evaluations are essential for instructors to understand students' perceptions of their learning gains and thus improve their teaching. As previously mentioned, many student evaluation responses have yet to be deeply analyzed as it is time-consuming [1]. Hence, the goal of the study is to provide a distinctive method for instructors to summarize students' survey responses to assess the perceptions of their learning gains, facilitating a more student-centered approach to learning. We see how NLP methods such as cluster analysis generate summaries and topics that students frequently discuss in their responses. In addition, we see how sentiment analysis measures students' tone and emotions in their responses. Further results from our research suggest that free responses may vary and contain more information compared to Likert-scale responses. By using cluster and sentiment analysis, we offer instructors a unique and user-friendly way to quickly and efficiently glean insights into students' perceptions of their learning gains, thus allowing instructors to take proactive measures that will improve their teaching. In particular, we expect the techniques discussed in this study to be applicable to other different subjects where there is a free response section in the course survey, as the essence of this study is to analyze students' free responses.

Exploring our first research question: *''To what extent is cluster and sentiment analysis effective for summarizing and analyzing Student Assessment of their Learning Gains (SALG) survey in algebra-based physics courses at a large Canadian research university?''*, we see how methods such as K-means and VADER may provide compelling insights for instructors, although there are some limitations, such as

TABLE 10. The keywords in the clusters and their corresponding topic for the 2020 SALG survey.

Cluster	Topic	Number of Student Responses	Keywords
0	Participation	296	useful, participation, participated, atmosphere, hall
1	Suggestions	276	resources, material, attitude, hard, lecture
2	Comments on the Subject	340	life, class, analyzing, evaluate, steps
3	Learning Methods	368	practice, future, learn, textbook, material
4	Learning Gains	307	study, formulas, problems, electric, world
5	Supports	307	lectures, session, study, resources, time
6	Course Structure	496	lecture, activities, discussions, demonstrations, atmosphere
7	Lecture & Textbook	538	communication, subject, experiment, resources, topics
8	General Comments	225	practice, interrogation, learned, readings, use
9	Helpfulness of Different Resources	184	help, practice, content, helped, understanding
10	Practicals	364	apply, needed, well, review, practical

TABLE 11. The keywords in the clusters and their corresponding topic for the 2019 SALG survey.

Cluster	Topic	Number of Student Responses	Keywords
0	Suggestions	256	taught, participate, felt, helped, information
1	Workload	436	always, prof, discouraged, lectures, notes
2	Understanding of Concepts	452	ta, seek, study, practical, participate
3	General Comments	277	things, catalytic, lecture, peers, activities
4	Comments on the Subject	382	quizzes, helped, always, school, world
5	Helpfulness of Different Resources	303	learn, videos, really, study, would
6	Problem Solving	257	able, think, answer, useful, reasoning
7	Learning Gains	328	example, like, simulations, think, activities
8	Lecture & Textbook	639	making, scientific, specific, ability, think
9	Learning Methods	420	helps, professor, lecture, material, catalytic
10	Graded Assignments & Test	476	explain, think, life, knowledge, topics
11	Participation	168	helped, visual, demonstrations, ideas, piazza

manually labelling the clusters for K-means and adequate accuracy when using VADER.

As for the sub-question: *"How effective are NLP summaries of free responses compared to the statistics from Likert-scale responses?"*, we see how information conveyed in both the cluster and sentiment analysis are fairly effective in being consistent with our impression that free responses may convey more information compared to Likert-scale responses. By extracting summaries for each cluster in each year of the SALG data, we see how we can give instructors an overview of student responses in a short amount of time. The comparison between the students' free responses and Likert-scale responses is measured through the median of the sentiment score and Likert-scale distribution, in which our results suggest that free and Likert-scale responses may vary in information.

Further questions to explore include finding a suitable method to choose the number of clusters and how to improve the accuracy of classifying negative responses. If one continues this research, one can apply silhouette analysis [33] to determine the optimal number of clusters to try to resolve the issue of choosing the number of clusters or use different clustering algorithms to cluster responses. Additionally, one can train a model using the Naive Bayes classifier [27] for higher classification accuracy or investigate different sentiment analysis algorithms. Lastly, one can strive to create a website or digital tool for instructors to input their course surveys and receive unique summaries from our cluster and sentiment analysis. This gives hope that in future years, student surveys can be used more frequently within a school year as instructors would be able to draw effective conclusions from the surveys in a shorter amount of time, advancing the approach of student-centered learning in PHE.

APPENDIX A
TABLES FOR CLUSTER ANALYSIS

See Tables 10 and 11.

APPENDIX B
CODE OF THIS PAPER

Link to GitHub Repository: https://bit.ly/3D9AYrp

REFERENCES

[1] S. Cunningham-Nelson, M. Laundon, and A. Cathcart, "Beyond satisfaction scores: Visualizing student comments for whole-of-course evaluation," *Assessment Eval. Higher Educ.*, vol. 46, pp. 1–16, Aug. 2020, doi: 10.1080/02602938.2020.1805409.

[2] B. Uttl, C. A. White, and D. W. Gonzalez, "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related," *Stud. Educ. Eval.*, vol. 54, pp. 22–42, Sep. 2017, doi: 10.1016/j.stueduc.2016.08.007.

[3] N. Grönberg, A. Knutas, T. Hynninen, and M. Hujala, "Palaute: An online text mining tool for analyzing written student course feedback," *IEEE Access*, vol. 9, pp. 134518–134529, 2021, doi: 10.1109/access.2021.3116425.

[4] E. Seymour, D. J. Wiese, A.-B. Hunter, and S. Daffinrud. *Creating a Better Mousetrap: On-Line Student Assessment of their Learning Gains*. Accessed: Nov. 13, 2022. [Online]. Available: https://salgsite.net/docs/SALGPaperPresentationAtACS.pdf

[5] S. Exarhos, "Anti-deficit framing of sociological physics education research," *Phys. Teacher*, vol. 58, no. 7, pp. 461–464, Oct. 2020, doi: 10.1119/10.0002061.

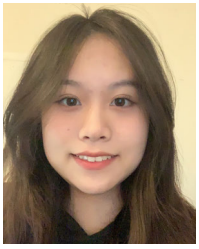
[6] P. M. Simpson and J. A. Siguaw, "Student evaluations of teaching: An exploratory study of the faculty response," *J. Marketing Educ.*, vol. 22, no. 3, pp. 199–213, Dec. 2000, doi: 10.1177/0273475300223004.

[7] P. Spooren, B. Brockx, and D. Mortelmans, "On the validity of student evaluation of teaching," *Rev. Educ. Res.*, vol. 83, no. 4, pp. 598–642, Dec. 2013, doi: 10.3102/0034654313496870.

[8] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.
[9] O. Almatrafi and A. Johri, "Improving MOOCs using information from discussion forums: An opinion summarization and suggestion mining approach," *IEEE Access*, vol. 10, pp. 15565–15573, 2022, doi: 10.1109/access.2022.3149271.
[10] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.
[11] S. Sayad, "K-means clustering," Introduction Data Sci., Tech. Rep., 2010. Accessed: Mar. 5, 2023. [Online]. Available: http://saedsayad.com/clustering_kmeans.htm
[12] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 199–205, Mar. 1982.
[13] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, May 2014, vol. 8, no. 1, pp. 216–225. Accessed: Jun. 4, 2022. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550
[14] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *Amer. J. Phys.*, vol. 66, no. 1, pp. 64–74, Jan. 1998, doi: 10.1119/1.18809.
[15] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 23, pp. 8410–8415, May 2014, doi: 10.1073/pnas.1319030111.
[16] F. Ornek, W. R. Robinson, and M. P. Haugan, "What makes physics difficult?" *Int. J. Emerg. Sci. Eng.*, vol. 3, no. 1, pp. 30–34, 2008.
[17] A. Bray and J. Williams, "Why is physics hard? Unpacking students' perceptions of physics," *J. Phys., Conf. Ser.*, vol. 1512, no. 1, Apr. 2020, Art. no. 012002.
[18] D. J. Exeter, S. Ameratunga, M. Ratima, S. Morton, M. Dickson, D. Hsu, and R. Jackson, "Student engagement in very large classes: The teachers' perspective," *Stud. Higher Educ.*, vol. 35, no. 7, pp. 761–775, Nov. 2010, doi: 10.1080/03075070903545058.
[19] G. Brauwers and F. Frasincar, "A survey on aspect-based sentiment classification," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–37, Apr. 2023, doi: 10.1145/3503044.
[20] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 6, p. e1332, Nov. 2019, doi: 10.1002/widm.1332.
[21] M. Hujala, A. Knutas, T. Hynninen, and H. Arminen, "Improving the quality of teaching by utilising written student feedback: A streamlined process," *Comput. Educ.*, vol. 157, Nov. 2020, Art. no. 103965, doi: 10.1016/j.compedu.2020.103965.
[22] K. Schouten, F. Frasincar, R. Dekker, and M. Riezebos, "Heracles: A framework for developing and evaluating text mining algorithms," *Expert Syst. Appl.*, vol. 127, pp. 68–84, Aug. 2019, doi: 10.1016/j.eswa.2019.03.005.
[23] S. Delgado, F. Morán, J. C. S. José, and D. Burgos, "Analysis of students' behavior through user clustering in online learning settings, based on self organizing maps neural networks," *IEEE Access*, vol. 9, pp. 132592–132608, 2021, doi: 10.1109/ACCESS.2021.3115024.
[24] K. Z. Aung and N. N. Myo, "Sentiment analysis of students' comment using lexicon based approach," in *Proc. IEEE/ACIS 16th Int. Conf. Comput. Inf. Sci. (ICIS)*, Wuhan, China, May 2017, pp. 149–154, doi: 10.1109/ICIS.2017.7959985.
[25] A. S. Alblawi and A. A. Alhamed, "Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, NLP and analytics," in *Proc. IEEE Conf. Big Data Anal. (ICBDA)*, Kuching, Malaysia, Nov. 2017, pp. 124–129, doi: 10.1109/ICBDAA.2017.8284118.
[26] N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Sentiment analysis: Towards a tool for analysing real-time students feedback," in *Proc. IEEE 26th Int. Conf. Tools With Artif. Intell.*, Limassol, Cyprus, Nov. 2014, pp. 419–423, doi: 10.1109/ICTAI.2014.70.
[27] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python*. Beijing, China: O'Reilly, 2009.
[28] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2758–2765, Mar. 2011, doi: 10.1016/j.eswa.2010.08.066.
[29] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, Aug. 1987, doi: 10.1016/0169-7439(87)80084-9.
[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008. Accessed: Jun. 4, 2022. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html
[31] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, p. e2, Oct. 2016, doi: 10.23915/distill.00002.
[32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Jan. 1987, doi: 10.1016/0377-0427(87)90125-7.
[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[34] K. Leung, "Micro, macro & weighted averages of F1 score, clearly explained," Medium, Tech. Rep., Jun. 2022. Accessed: Nov. 13, 2022. [Online]. Available: https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f#2f35
[35] F. Taverna, M. Neumann, and M. French. (Jul. 10, 2019). *Innovating the Large Class Experience: Teaming Up! To Enhance Learning*. Accessed: Jun. 4, 2022. [Online]. Available: https://www.learntechlib.org and https://www.learntechlib.org/p/210299/

HONGZIP KIM was born in Vancouver, Canada, in 2003. He received the Associate's Diploma degree in piano performance (ARCT) from the Royal Conservatory of Music (RCM), in 2021. He is currently pursuing the H.B.Sc. degree in computer science and statistics with the University of Toronto, and he is specializing in artificial intelligence (AI) and computational linguistics and natural language processing (NLP).

His research interests include data science, machine learning, AI, and NLP. He is particularly interested in how education can foster a more inclusive and welcoming learning environment by analyzing students' free responses.



GETING QIN was born in Zhengzhou, China, in 2003. She is currently pursuing the H.B.Sc. degree in physics with the University of Toronto, ON, Canada. She specializes in quantum information and quantum computing.

From May 2023 to August 2023, she was a Student Researcher in atmospheric physics with the University of Toronto funded by the Natural Sciences and Engineering Research Council of Canada. Her current research interests include natural language processing, artificial intelligence, quantum machine learning, near-term quantum devices, and physics education.