Sentiment and thematic analysis of faculty responses: Transition to online learning

Colin Green, Eric Brewe, and Jillian Mellen
Physics Department, Drexel University, Philadelphia, Pennsylvania 19104, USA

Adrienne Traxler
Department of Science Education, University of Copenhagen,
Universitetsparken 5, København Ø, Denmark

Sarah Scanlin
Electrical Engineering, Drexel University, 3181 Chestnut Street, Philadelphia, Pennsylvania 19129, USA

(Received 5 January 2023; accepted 18 April 2024; published 12 June 2024)

This project aims to understand physics faculty responses to transitioning to online teaching during the COVID-19 pandemic. We surveyed 662 physics faculty from the United States following the Spring 2020 term; of these, 258 completed a follow-up survey after the Fall 2020 term. We used natural language processing to measure the sentiment scores of 364 Spring 2020 responses and another 134 Fall 2020 responses of physics faculty who completed an optional written prompt. Additionally, we determined the change in sentiment scores of the 100 individuals who responded to both surveys. These sentiment scores measured between −1 and 1 for completely negative and completely positive, respectively. Sentiment scores after Spring 2020 were slightly positive with a median value of 0.2347. The distribution of sentiment changes was approximately normally distributed with a mean centered near zero. Analysis suggests the average sentiment did not change from the initial to follow-up surveys. To identify major topics within the responses for both surveys, latent Dirichlet allocation analysis was applied to the data. The topic distribution for the initial survey is given as course modifications and technology, negative aspects of the transition—primarily with labs and cheating, exam and evaluation difficulties, and difficulties with student understanding. The topics were noticeably different in the follow-up survey with differences between Fall and Spring, cooperative learning strategies, strategies that worked in the remote space, and benefits of in-person labs.

DOI: 10.1103/PhysRevPhysEducRes.20.010151

I. INTRODUCTION

Everyone had to scramble to comply with new guidelines and reduce person-to-person contact. It became apparent that due to many factors, education had to deal with unique challenges during this pandemic. The speed at which nearly all educators had to engage with virtual learning was without much warning or transition time. Schools were going virtual, but there was very little room for training, professional development, or in some cases, securing the necessary equipment.

The transition to virtual learning was viewed as a tremendous shakeup of the established paradigm. Lewin [1] describes the process of change as happening in three stages. First is an "unfreezing" event that upends the current state, then change in a large organization such as education is possible. Instructional change is not always seen as an easy or desirable outcome, particularly embedded in the context of a pandemic. Investigating the sentiment, or how positive or negative responses are, will allow us to evaluate and quantify how instructors felt as they made the transition. Further, by looking at how sentiment changed over time, we get a better sense of how faculty became more comfortable or enthusiastic or less threatened by these changes.

Sentiment analysis is a machine learning tool utilized within the realm of natural language processing. Sentiment analysis allows us to gauge the relative positive or negative sentiment within a segment of text by computing a sentiment score. This sentiment score is typically between −1 and 1, where negative numbers are negative sentiment and positive numbers correspond to positive sentiment. We chose sentiment analysis as an analytic tool because it is able to quickly analyze large datasets that would take far too long by hand. Additionally, with a sentiment analysis algorithm, inherent biases can be moderately to strongly eliminated. Such biases are often exposed when a coder or reader of a passage imposes their own beliefs onto the text and adds error to the sentiment value. Sentiment analysis has become a more commonly used tool within other disciplines as well as a useful tool to analyze social media. There is little work within physics education research that uses sentiment analysis, so in this paper, we hope to identify areas where such techniques can be used.

II. LITERATURE REVIEW

A. Transitioning teaching during COVID

Evidence-based instructional practices that use extensive peer-peer interactions pose unique challenges when attempting to facilitate online instruction. Physics instructors who engage in evidence-based practices have been required to adjust their strategies in order to present them in the online space [2]. Surveys taken shortly after a number of institutions transitioned to online learning showed instructors had increased levels of anxiety compared to face-to-face instruction that they were more familiar with [3]. This research additionally showed that instructors relied most heavily upon other instructors for resources during this time. Other studies confirmed anecdotal beliefs that instructors struggled with most aspects of their work such as work satisfaction, grant writing, authorship, among others [4].

Looking at student outcomes during the COVID transition showed that first year undergraduates believed they struggled more than those in the upper level courses. The students who self-reported higher levels of organization lead to higher perceived learning outcomes [5]. Several attempts were made to institute new online learning paradigms to varying degrees of success. Researchers found many institutions, particularly institutions with diverse student populations, underestimated the degree of access that students had to the technology required for virtual learning [6]. COVID related research and retrospectives are beginning to emerge [7], however, much of this work is focused on students and their experiences. There is currently less research, and an even smaller amount in physics, being done on faculty experiences over the course of the pandemic.

B. Natural language processing

Chowdhury described natural language processing (NLP) as the “aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks” [8]. NLP tools are constantly being created and adapted to include new tasks and implementations. For example, text to speech has been widely implemented not only in everyday usage but also offers quality of life improvements for those unable to read. Commonly used machine translators use NLP to make translations more accurate, particularly when words have multiple meanings and the intent of word usage must be determined by looking at the semantic qualities of the word(s) and their neighbors [9]. Another large attractor of NLP is to mine social media, a method gaining widespread popularity in marketing and advertisements. Some notable applications provide opportunities for lifesaving intervention in using NLP as a method to screen for those at risk of suicide [10] and other adverse effects of social media [11]. There is no shortage of techniques that fall under the umbrella of NLP, and there is not always an obvious technique to obtain a desired result.

TABLE I. Four principles of quantitative text analysis described by Grimmer and Stewart [12].

1.	All quantitative models of language are wrong—but some are useful.
2.	Quantitative methods for text amplify resources and augment humans.
3.	There is no globally best method for automated text analysis.
4.	Validate, Validate, Validate.

With any machine learning approach, it is imperative that we understand what machine learning can, and cannot, tell us. Grimmer and Stewart [12] offer their four principles of text analysis (Table I). We adopt these principles in our investigation as a framework in which to orient our questions, methods, and analysis. In this paper, we introduce two particular NLP-based approaches, sentiment analysis and thematic analysis, for text-based analysis.

C. Sentiment analysis

There is little prior work within physics education research that uses sentiment analysis. Kelley *et al.* [13] used it to code student forum posts as positive or negative and to estimate the prevalence of different emotions over the semester. Gavrin [14] used the technique to code use of emotion words in student forum posts and end-of-semester evaluation comments in the pandemic spring of 2020. With online text data becoming more common in many educational areas including physics, we anticipate these techniques may be of growing interest to the PER community.

Sentiment analysis has two primary branches or approach methods: lexicon based and machine learning based approaches. The lexicon based approach uses a predetermined lexicon, or word and phrase bank, with corresponding sentiment scores. The machine learning approach uses methods such as support vector machine, Naive Bayes [15], and neural networks [16], among others. These approaches are useful in their own rights and provide an option not available to the lexicon-based methodologies in which the machine learning approaches are entirely self-contained. Such self-contained machine learning does not draw, or is less dependent, on external and predetermined sentiment statistics or evaluations. However, because lexicon-based algorithms are pretrained, they support smaller datasets. Lexicon sentiment analysis programs typically handle preprocessing of data, making them friendlier to use, and have been shown to trend with Likert scale results in student course evaluations [17]. For these reasons, we will be using a lexicon-based approach for our analysis.

D. Thematic analysis

An additional gauge of how physics instructors experienced the instructional change would be to extract the major themes that were being discussed in the responses of those instructors. Thematic analysis can illuminate the topics that were of the most concern to instructors during two phases of the instructional shift: in the Spring of 2020 and the following semester or quarter in the Fall of 2020. By using thematic analysis at two different time points, we can also look for shifts in the themes.

Latent Dirichlet allocation (LDA) [18] is an unsupervised tool for doing thematic analysis. The approach for LDA is to take all of the words in a "document"—in this case a single response—and plot the position of the documents in a so-called "Word Space." This word space is a vector space that is spanned by all of the words in the collection of documents (commonly referred to as a corpus). In this analysis, each document forms a position in the corpus space. The LDA algorithm then begins to identify clusters of documents in the space and develops parameter-based optimized clusters which form the collection of documents in the "topics" [19].

We use sentiment and thematic analysis tools to extract underlying information from the survey response text. Sentiment analysis identifies how each physics instructor was feeling during the online transition in 2020, while LDA determines what overarching themes were present in the responses.

III. RESEARCH QUESTIONS

RQ1 Did the favorability of physics instructors change from Spring to Fall?

RQ2 What themes were present in the Spring and Fall responses?

RQ3 How did the themes change as physics instructors had more time to develop their strategies?

IV. METHODS

A. Data collection

These data were collected as part of a larger study on the transition to online learning [3]. The data were collected from a survey that was sent to the 10 largest universities in each of the 50 states as well as the District of Colombia, Puerto Rico, and the U.S. Virgin Islands (20 universities were chosen in the following states: California, Florida, New York, Pennsylvania, and Texas), yielding a list of approximately 600 schools. We first confirmed that each university had transitioned to online teaching. Then we used web scraping to collect email addresses, resulting in a list of approximately 14,000 email addresses in total. No incentives were given for responses. After filtering responses to include those who were teaching physics during Spring 2020 and completed more than 50% of the survey, we were left with 662 on the initial survey. The initial survey contained 38 questions in total, with the 37th question (Q37) being a free response question: "We are interested in your experiences with teaching your physics class online. Feel free to describe how you went about transitioning the class, any lessons you have learned and anything you think we should understand." We received 364 responses to this prompt on the initial survey (55% of all respondents). We followed up with the same 662 respondents following the fall 2020 term. The follow-up survey resulted in 258 responses that met the requirements of the data collection; of these 258 responses, 134 responded to the open-ended prompt (52% of all respondents).

There was a wide range of response sizes, with the largest response consisting of 2884 words, while the shortest response was only 1 word ("horrible").

Matching data to include only the participants who responded to the open-ended prompt in both the initial and follow-up surveys resulted in 100 matched pair responses. The paired responses were used in the sentiment analysis, while all responses were used in the thematic analysis. All matching and analysis were done with anonymous identification numbers created at the beginning of data collection.

B. Demographics

The demographic breakdown and gender identity of all the respondents came out as follows: 71% identified as male, 22% identified as female, 1.1% identified with a nonbinary gender option, and the remaining 2.9% declined to state. The racial or ethnic background of the respondents was self-reported as 74% indicated they were White, 9.8% Asian, 4.2% indicated Hispanic, Latino/x or of Spanish origin; 1.2% Black or African American, and 1.1% North African or Middle Eastern. Two individuals identified as American Indian or Alaska Native and two identified as Native Hawaiian or Pacific Islander. An additional 2.3% indicated their racial or ethnic background as "Other" and 5.3% preferred not to answer.

Based on American Institute of Physics demographic data, the demographic representation of physics faculty in the United States is 20% women, 79% White, 14.2% Asian, 3.2% Hispanic, 2.1% African American, and 1.2% other. [20] We do not wish to compare these numbers too strongly as the questions used by our survey and AIP are different.

C. Sentiment analysis using VADER

To analyze the written responses, VADER, a machine learning algorithm within the natural language tool kit (nltk) [21] was used in the Python coding environment. VADER (Valence Aware Dictionary for Sentiment Reasoning) is an open source lexicon and rule-based sentiment analyzer used commonly in social media analysis [22]. VADER estimates the compound (or total) score that ranges from -1 (completely negative) to $+1$ (completely positive) to represent the sentiment of a piece of text, with scores between -0.05 and 0.05 considered to be "neutral." The VADER algorithm relies on a constantly updating imported lexicon, or word and phrase base, that maps the words, phrases, and special characters to sentiment scores.

This type of machine learning algorithm could fall within the category of supervised learning. Supervised learning refers to algorithms that have a training set with known quantities. With a program like VADER, the training is done externally by the designers of the program. This algorithm then takes in new text—in this case, our responses—and maps the new text using the training data to give us a compound score for the response as a whole. The training data used for VADER are obtained from the lead researchers in the development of the algorithm. They describe their training set as being empirically verified by independent human judges [22]. The algorithm itself takes the input words and breaks them into individual units or tokens. There are internal mechanisms within the algorithm for text preparation such as tokenization (separating all the words or units in the text), removing capitals, removing symbols, and punctuation among others. The sentiments of the tokens are then calculated and averaged to create a compound sentiment score for the whole text.

VADER was chosen for its previous use within physics education research, ease of use, reliability, and human verified gold standard lexicon to draw from [23]. The lexicon-based approach with VADER allows us to calculate sentiment without requiring segmenting the data into a training and testing set. VADER is based off a lexicon that has its values input by human researchers in the language-based fields. This lexicon is imported and updated by external researchers and not the authors of this paper. VADER has been shown to trend with Likert scale results in student course evaluations [17].

Our team tried other sentiment analysis tools, such as Textblob [24]. The initial responses were appropriately similar for us to move forward with VADER with confidence as it is consistent with other open-sourced tools.

A primary concern when dealing with sentiment analysis and the English language is the ability for n-grams (n-number of words that, for the purposes of analysis, should be treated as a single unit) to skew our results. An example of such a phrase could be "did not enjoy," a phrase that we expect to score moderately to distinctly negative. The most basic sentiment analysis would see us assigning a sentiment score to each word in that phrase individually, followed by taking the sum of each score to give the total sentiment of the phrase. Using VADER itself to accomplish this, "did" and "not" resulted in compound scores of 0.0 while "enjoy" returned a compound score of 0.4939. The error in this approach is apparent: the measured sentiment is positive, contradicting expected results. This phrase is perfect for analysis because it includes a bigram, a pair of words that the algorithm should take together. The word not changes the sentiment of the following word(s) and we need the algorithm to be responsive to such phrases. To check VADER operates on these phrases, we ran the same phrase "did not enjoy" through VADER's sentiment calculator. The result of the analysis by VADER resulted in a compound score of -0.3875. This score verifies that VADER does the searching for these n-grams within its operation and allows for quicker, more accurate sentiment scores [22].

Several responses exemplify their respective compound sentiment scores in order to verify VADER was working properly.

"Testing is problematic in a large class. I have no way to proctor that many online students. Cheating is rampant. Also, I teach astronomy; labs are difficult online. There are also no options for real-time observations with telescopes online. You cannot recreate the experience of being under the night sky." This was recognized as an overall negative response with a compound score of -0.9081.

"In Spring 2020 we went online in the middle of the semester. We had to improvise on a dime. Spring break started normally but was extended ine [one, sic] week to give us (and the students) to get ready. Fall 2020 was planned from the beginning to be online only. I was more comfortable than most because I had taught an online asynchronous Conceptual Physics course (that I developed) for many years, and so knew many approaches and routines that were valuable. Going online but synchronous was actually much easier than asynchronous. But I will rejoice the day when we can go back in the classroom. The students do perform better in a Asynchronous format than asynchronous [sic] because of two major factors: the regularity of the meetings helps keep them on task and, to my surprise, the fact that they can comment and participate by asking question without being "seen" (camera off). It reduces their unwillingness to be recognized in a classroom setting." This response received a positive compound score of 0.9845.

"As a TA I didn't have much control. We recorded the expirements [sic] and had students watch them and go about their usual work." This response received a very neutral score of 0.0.

1. Comparing sentiment scores

The compound scores from the initial and follow-up surveys were tested as matched pairs using a Bayes factors t test approach. There are two primary ways that Bayes factors are desirable in this project. First, the odds ratios using a Bayesian framework are more easily interpretable than p values, e.g., an odds ratio of 2∶1 says the evidence is roughly twice as strong for one model as the other and is seen as fairly weak evidence in favor of the first

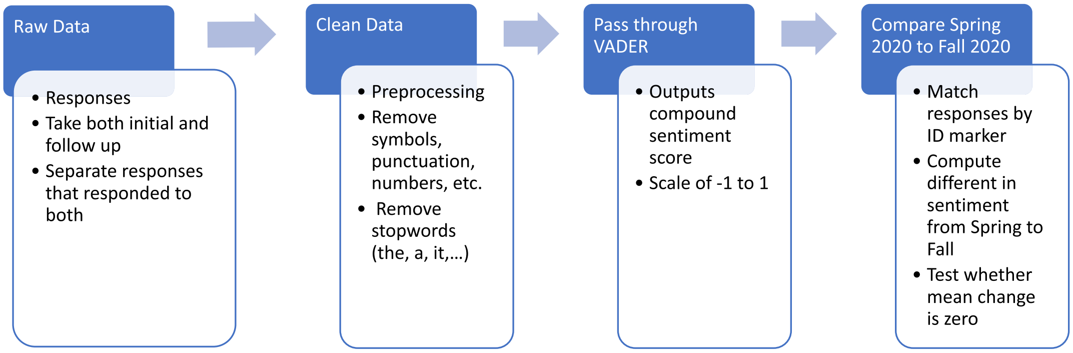

Raw Data
• Responses
• Take both initial and follow up
• Separate responses that responded to both
Clean Data
• Preprocessing
• Remove symbols, punctuation, numbers, etc.
• Remove stopwords (the, a, it,…)
Pass through VADER
• Outputs compound sentiment score
• Scale of -1 to 1
Compare Spring 2020 to Fall 2020
• Match responses by ID marker
• Compute different in sentiment from Spring to Fall
• Test whether mean change is zero

FIG. 1. Simplified overview of the sentiment analysis process using VADER.

model. Second, using a Bayes factor approach can provide evidence in favor of a null model, which a standard t test cannot do [25]. This approach thus provides a similar result to standard matched pair t tests but can be used to claim more than the student's t test is able to.

Bayes factors are calculated in a two-part approach. First, Bayesian models for the null and alternative are estimated using Markov Chain Monte Carlo methods, and then an odds ratio between the null model (no mean differences) and the alternate model. Thus the odds ratio effectively quantifies the strength of evidence in support of the alternate model. These ratios are invertible so can also provide evidence in favor of the null model. In this project, we used the BayesFactor [26] package within the R statistical programming language [27] to determine support for a null hypothesis of no change in sentiment from initial to follow-up.

A simplified and concise pictorial representation of our process can be viewed in our sentiment analysis flowchart (Fig. 1).

D. LDA analysis

As mentioned, there are several parameters that guide the LDA algorithm. Three that are needed include the number of topics (K); a parameter, alpha, which represents how "mixed" the topics are within the documents; and finally, an initiation or seed value. The alpha parameter is the more abstract of the parameters and typically requires grid-searching to maximize. However, with smaller datasets, a slower self-learning LDA model can be run that searches itself for the optimal alpha value to use. Due to the size of our dataset, we opted to use this automatized alpha search LDA. The number of topics was searched for using a grid-search type approach. The value of K can be sought by looking at the stability of the LDA model at different K values. The stability of the model is defined by how the model looks at a specific K value when you vary the random seed. Stable models will have limited variation under random seed changes. It is very possible to create a better evaluated model with LDA for one K value over another by simply getting lucky with a specific random seed used [28].

We used a grid-search technique to evaluate the effectiveness of each model. For LDA, there are several ways of doing this, but one of the most common is to use the "coherence score." This score looks at the relative semantic similarity between the top words in each topic [29]. Coherence score is calculated through the GENSIM module [28] and is the standard metric for LDA topic modeling. One conceptualization of coherence score is to think of it as a measure of how clustered the items within the topics are to each other. A coherence score that is close to 1 would indicate that the documents or responses are using all the same words and would be extremely clustered. This would indicate that the algorithm has effectively identified these clusters. A coherence score near 0 indicates that there are no discernible groups of responses—all the responses are evenly spaced and you would be unable to identify any topics that connect more than one response. As of writing, there appears to be no consensus as to what constitutes a good or bad coherence score. Instead, we opt to maximize the coherence score based on the data given and take these measures to optimize different parameters such as K. We combine the coherence score search with other noncomputational methods such as increasing the amount of human input and validation to further increase the accuracy of the thematic analysis. A larger coherence score is generally associated with a more powerful model. Our coherence scores are lower on average than similar LDA studies [19]. This study and results from Syed and Spruit's paper on coherence scores [30] suggest that the average value of coherence score is directly correlated with the size of the text database used in the analysis. This goes some way toward explaining our specific range of coherence score values in Figs. 2 and 3. We will primarily be using the coherence score as a comparative tool between different K values, helping us choose the most appropriate topic number.

The grid search for K was conducted by running an LDA model with alpha="auto" for K values 3 through 10. For each K value, the random seed was varied 10 times. We plotted the distribution of the coherence scores for each of the topic number values and runs.

The grid search for the initial data in Fig. 2 shows that the average coherence score appears to drop off as the number

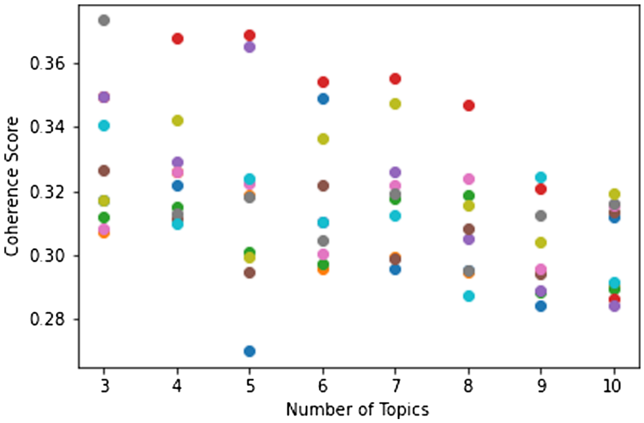


FIG. 2. Number of topics (K) search for initial dataset.

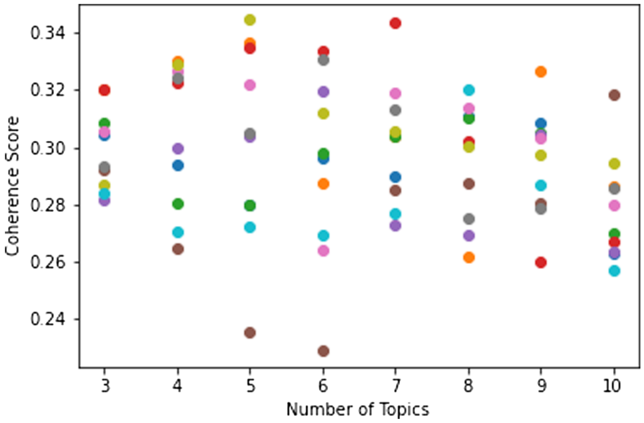


FIG. 3. Number of topics (K) search for follow-up dataset.

of topics gets larger. Unfortunately, there is not a large amount of differentiation between the coherence score distributions for $K = 3$ or 4. Using the coherence scores to further refine the optimal K value is unrealistic and so we conducted a manual check, running the LDA algorithm with both $K = 3$ and $K = 4$. From each of those models, the responses that were most heavily associated with each single topic were examined by six PER researchers familiar with the project to identify the general coherence between those topics. From those results, we concluded the best fit for the number of topics in the initial responses was 4. This is a significantly more manual approach to topic number identification that is often utilized. Machine learning examples often have massive datasets and therefore can often determine these properties without much, or any, manual evaluation. That was not the case for us, as we were required to intervene, particularly in the follow-up where we had the smallest dataset. The manual approach we employ does not necessarily reduce the validity of the results, it indicates the computer needed more human input than it may otherwise with larger datasets.

The grid-search technique was less distinct for the follow-up responses (Fig. 3), likely due to the smaller dataset. While specific topic numbers had larger coherence scores than others, the lack of stability of those same models poses issues with identifying them as the optimal K value. These difficulties considered, we decided to take a more manual approach to the follow-up data, similar to the approach to identify whether $K = 3$ or 4 was the appropriate choice for the initial data. This slightly different approach from the initial survey was done after discovering the coherence scores were not giving any distinct cutoffs to narrow our focus on the value of K. Using the same methodology of using multiple manual coders, we looked at the results of separate LDA runs with $3 < K < 10$. The larger numbers of topics mixed too aggressively, removing the uniqueness and individuality of each topic, and making any topic identification not feasible. After manually looking at $K > 5$ runs of thematic analysis, the manual topic identification became impossible and indicated to us that the size of the dataset was suggesting these K values were too large. Using the manual topical identification, $K = 4$ was found to be the best fit for the follow-up responses.

It is interesting to note that there is no intrinsic theoretical reason for the number of themes to be the same in the initial and follow-up responses. The two responses having different numbers of themes could be a perfectly reasonable result of the analysis, particularly because the difference in the size of the datasets. While the fact that the two K values are the same is not required, nor should it be viewed as striking or overly meaningful. Rather, we wish to highlight the need for analytical analysis of these topic or theme numbers, they should not be assumed.

1. Data cleaning for LDA

In order to apply LDA thematic analysis to our responses, the input to the LDA must be a “bag of words matrix.” This is a matrix that has each response placed on the rows of the matrix; the columns of the bag of words matrix are then each item in the corpus (that is to say all words that are in the collection of responses). The bag of words matrix is then filled in with entries of 1 or 0 depending on whether the specific column word was present in the specific response.

The usefulness of LDA is often marred by the additional, and in LDA’s case unimportant, extraneous parts of text. To increase the effectiveness of LDA, the first step in the analysis was to apply several data cleaning initiatives. This preprocessing of the responses was done primarily with the use of Python tools including PANDAS [31,32], GENSIM [28], and the NLTK [21] packages. The first step was to begin with a dataframe of the responses to question 37. We wanted to search for common bigram phrases; these are words that show up and make sense only when viewed as a pair or bi-gram. These common phrases, such as “face-to-face” are then combined into a single token “face_face” and analyzed by LDA as a single token [19]. Some examples of

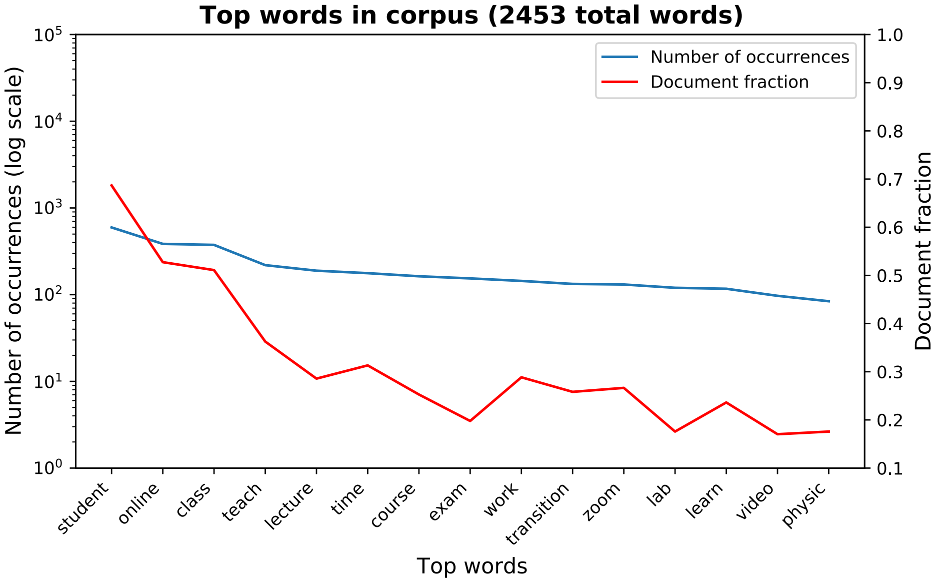


FIG. 4. Collection of word distribution in the initial responses following data cleaning.

bigrams found were "breakout_room," "small_group," "real_time," "upper_level," and more. The responses were then run through several cleaning steps to remove punctuation, numbers and symbols, and to lowercase all words. Following this, we needed to remove stopwords, which are words that bloat the vector space of LDA analysis and cloud the results. Examples of stopwords are "the," "it," and "to."

With the data primarily cleaned, the responses must be tokenized (separating each individual word in each response). The last step in the data cleaning process was to lemmatize each word; this checks for words that have the same base word but show up in the corpus again because of changes due to part of speech, pluralization, etc. An example of lemmatization would be the program taking in the words "study" "studying" "studies" and reducing them all to the same word. Lemmatizing the words reduces the chances that LDA will become distorted by these differences [19]. There is some evidence that lemmatizing words has an effect on the results of machine learning algorithms with languages that have specifically rich morphologies such as Russian [33]. However, this research does point to this being impactful with languages such as English.

Following the preprocessing, we can look at the distribution of words within each of the initial responses shown in Fig. 4.

When utilizing LDA analysis, a frustration can occur where specific words that are common among a large portion of the documents can obscure the differentiation of the intrinsic topics. To address this, we began by removing all words that appear in 55% or more of the documents. As seen in Fig. 4, this removes the word "student." 55% was chosen based off other LDA research [19] and based on removing words that are in a "majority" of the responses. This was a starting point for removing these dominant words that can overwhelm or overpower the analysis. This 55% was a preliminary majority removal that was implemented with the belief there would be additional words that would need removal. This additional removal is *ad hoc* and used in conjunction with trial run coherence score grid searches and top word results (see Table II). These trial runs of LDA began to indicate that certain words were appearing in every topic and "absorbing" too much of the topics. In an

TABLE II. A selection of the top words for the different topics identified by the LDA analysis.

Theme number	Top words with weights
Spring 1	lecture: 0.031, zoom: 0.023, course: 0.022, work: 0.018, transition: 0.015
Spring 2	exam: 0.032, transition: 0.019, learn: 0.017, course: 0.016, work: 0.015
Spring 3	course: 0.013, face_face: 0.012, work: 0.012, lecture: 0.011, content: 0.011
Spring 4	week: 0.023, lecture: 0.019, video: 0.015, lab: 0.015, learn: 0.014
Fall 1	learn: 0.041, work: 0.040, remote: 0.028, time: 0.027, video: 0.023
Fall 2	taught: 0.034, person: 0.032, lecture: 0.027, grade: 0.023, format: 0.022
Fall 3	semester: 0.034, lab: 0.026, go: 0.023, well: 0.022, test: 0.020
Fall 4	time: 0.032, learn: 0.030, like: 0.030, work: 0.021, question: 0.021t

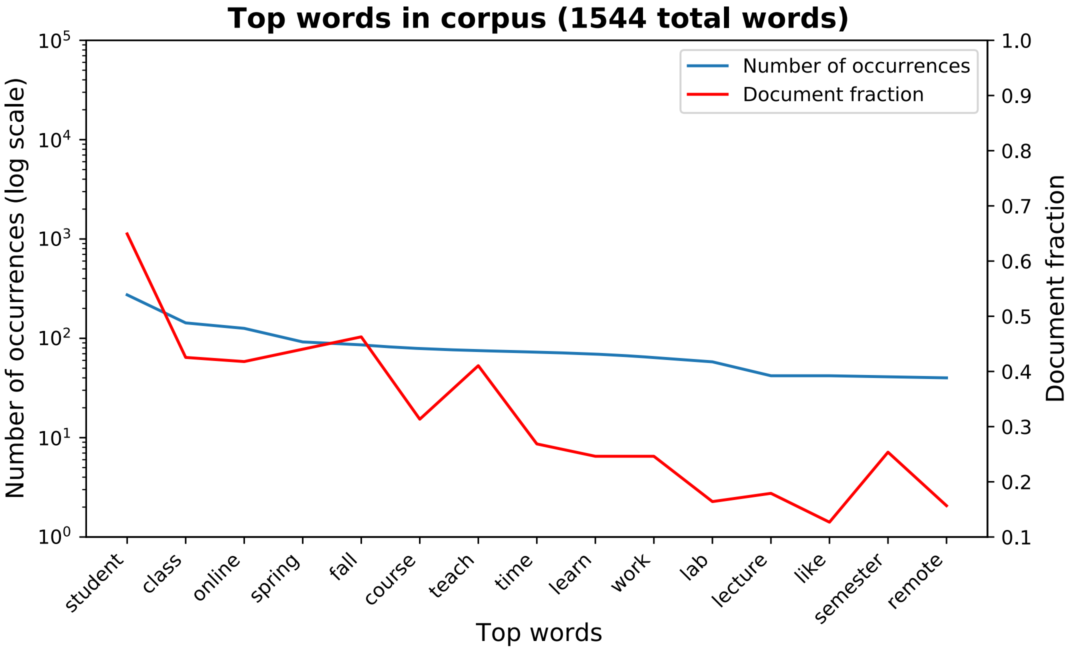


FIG. 5. Collection of word distribution in the follow-up responses following data cleaning.

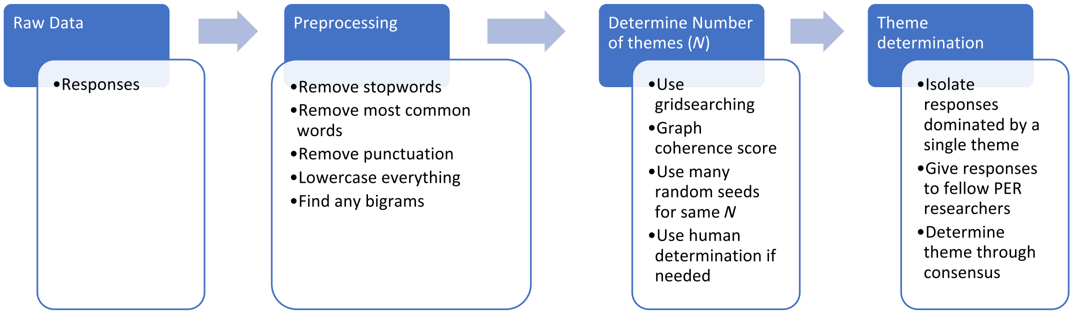
effort to counter this, we removed the following words: “class,” “online,” “teach,” and “time.” These words are common in the responses but do not function to identify the individual topics.

We repeated the same data cleaning procedure for the follow-up responses. We begun by removing any words that appear in 55% or more of the responses. This resulted in the removal of the word “student.” To identify if there were any further words that needed removing, the LDA analysis was run several times; no such patterns of topic merging that showed themselves in the initial responses were seen in the follow-up analysis, so no additional words were deemed necessary to remove. A word distribution for the follow-up responses can be seen in Fig. 5.

2. *Identifying topics*

The output of LDA analysis is a dataframe that gives the topic distribution of each response. LDA is not able to qualify what each topic represents. The first step toward identifying the topics required determining the responses that were dominated by a single topic. LDA analysis results in each response being described by a vector of length equal to the number of topics. The components of this vector are the proportion of the response that lies within that specific topic. An example could show one response being given by the vector (0.9, 0.05, 0, 0.05); this corresponds to a response that is 90% within topic 1, 5% topic 2, 0% topic 3, and 5% topic 4. We identified all of the responses that were dominated by a single topic; we set a limit of a response having a single topic value of greater than 90%. 90% was chosen as we believe this to be an appropriate cutoff for topic dominance, while maintaining enough samples to adequately investigate by hand. Those selected responses were sent out to six researchers who were instructed to read through each of the responses in a given topic and summarize the overarching theme of those responses. The follow-up responses were analyzed in an identical way.

Similar to sentiment analysis, we offer an overview of the process of thematic analysis in Fig. 6.


Raw Data

- Responses

Preprocessing

- Remove stopwords
- Remove most common words
- Remove punctuation
- Lowercase everything
- Find any bigrams

Determine Number of themes (*N*)

- Use gridsearching
- Graph coherence score
- Use many random seeds for same *N*
- Use human determination if needed

Theme determination

- Isolate responses dominated by a single theme
- Give responses to fellow PER researchers
- Determine theme through consensus

FIG. 6. Flowchart showing the general process of the application of thematic analysis of the data.

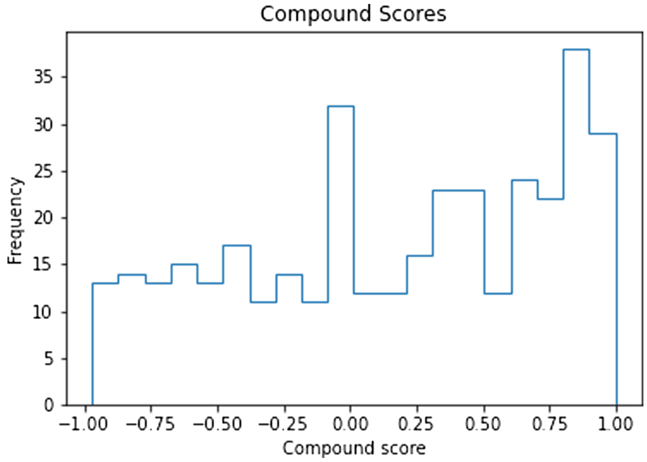


FIG. 7. Compound sentiment score for all initial responses.

V. DATA AND RESULTS

A. Sentiment of initial and follow-up survey

We applied sentiment analysis to all 364 responses from the initial survey and all 134 responses from the follow-up survey. The compound score results can be seen in Fig. 7; there is little evidence of a dramatic skew or significant pattern. The median sentiment score in the initial survey was 0.2347.

The distribution of sentiment scores on the follow-up survey can be seen in Fig. 8. There is a similar shape to that of the initial survey results. The median sentiment score in the follow-up survey was 0.2960.

B. Matched pair analysis

The 100 individuals who responded to both the initial and follow-up surveys had their responses matched using anonymous identifiers (ID). The difference in compound sentiment score was calculated for each ID.

Before analyzing the change in sentiment, we look at how representative of a sample the 100 individuals in the matched pair analysis are. To do this, we take our 364 initial responses and separate them by whether they responded to the follow-up or only the initial survey. We calculate the sentiment scores for these two groups and display the corresponding histograms in Fig. 9. This allows us to get a sense of whether the individuals in the matched pair analysis are drastically different from the other individuals in the survey. Should the histograms in Fig. 9 be significantly different, which implies there could be significant bias in the matched pair analysis. The results of this analysis do show the two groups being fairly similar, with no difference significant enough for us to not move forward with the matched pair analysis.

The histogram of these differences is shown in Fig. 10, where the change in sentiment scores appears to be fairly normal centered just above zero.

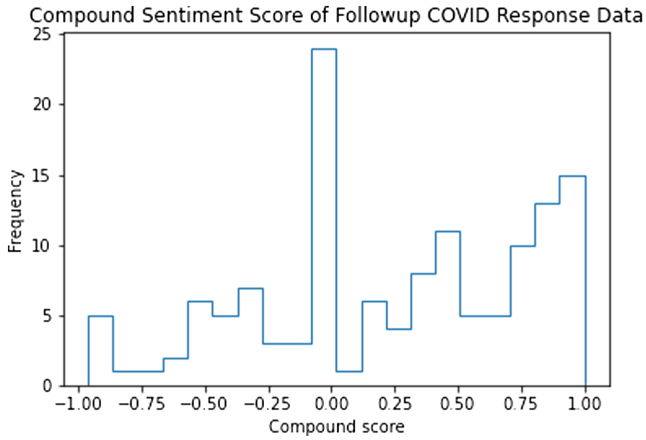


FIG. 8. Compound sentiment score for all follow-up responses.

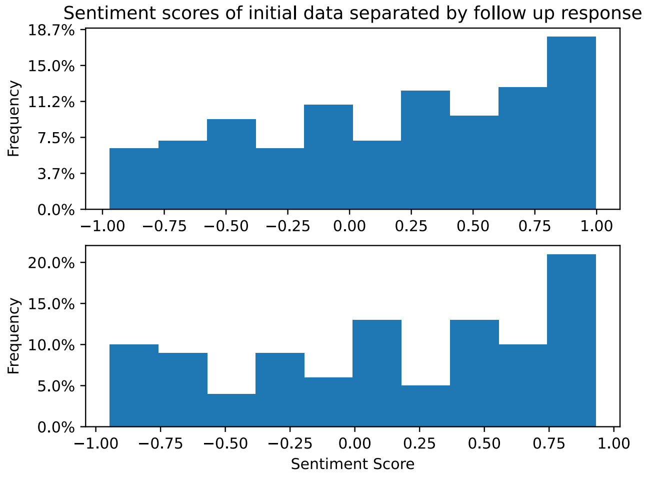


FIG. 9. A look at the sentiment distributions for the Spring 2020 responses from individuals who only responded to the initial survey (top) and the responses from those who responded to both the initial and final survey (bottom).

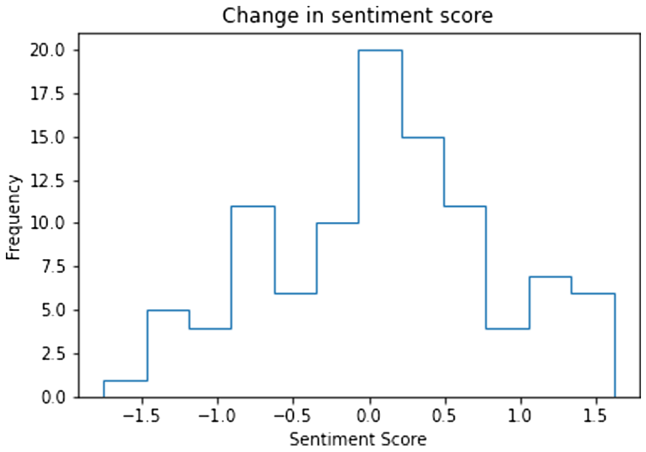


FIG. 10. Change in sentiment scores for 100 matched written responses.

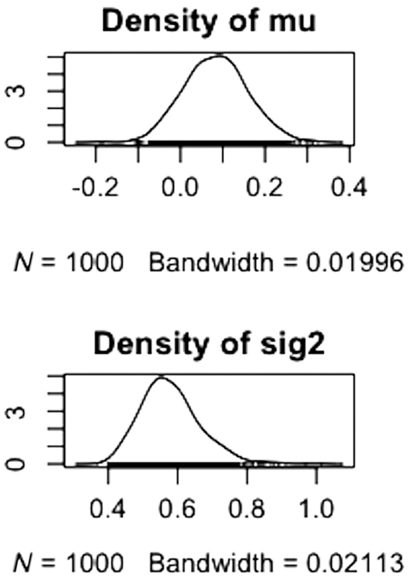


FIG. 11. Bayesian analysis of the change in sentiment score for matched responses.

The Bayes factor analysis, with 1000 iterations, found the difference in sentiment from initial to follow-up resulted in a mean change of 0.08375 with a standard deviation of 0.76112. These confirm the visual interpretation of Fig. 10. The Bayes factor analysis gives a density of mu, the mean sentiment score, plot shown in Fig. 11; this shows a normal distribution centered at approximately 0.1.

Calculating the Bayes Factor for the odds in favor of the null hypothesis (no change in sentiment) resulted in a Bayes Factor of 4.9∶1. This means that the data provide evidence that it is about 5 times more likely there was no average change in sentiment than it is there was a change. This is seen as weak-moderate evidence in favor of the null hypothesis, meaning that we did not measure a difference in sentiment from Spring 2020 to Winter 2021 [34].

C. Cases with large changes

To get a better look into some of the "blackbox" parts of VADER, we pull out two of the cases that show a large change in sentiment from Spring to Fall. These cases are a snapshot of the results chosen for their sentiment values and overall length, as we did not want each response to be a page long.

Our first case shown below is one instance of a positive shift in sentiment score. The response to the survey in the Spring of 2020 had a compound score of −0.9177 and was as follows:

> Our university required that our classes meet synchronously; remote, but not "fully on-line" The worst aspect was the rampant cheating (classmates & Google & Chegg) by some students.

this quote had a generally negative sentiment score, likely coming from the "rampant cheating" and describing it as the "worst" aspect.

We can compare this to the same individuals response during the follow-up survey in the Fall of 2020 with a score of 0.4902:

> Fall 2020 we knew in advance about the pandemic. I made all our intro labs as F2F with one (1) remote section late Friday afternoon (for anyone who didn't go to a F2F section that week) … so much better than trying to "blend" a lab.

here we see a more positive sentiment coming through. There is little in this response that pulls the sentiment negative, and there are several aspects that likely pushed the score toward the positive such as the description of this new lab technique as "so much better" and knowing about the pandemic in advance.

The second case we will look at shows the opposite trend to the first, an individual whose response to the survey went from a positively scored sentiment to a negative sentiment. These individuals response to the survey in the Spring of 2020 had a compound score of 0.8481 and response:

> I made my own video recordings of lectures. The exams changed from in-class, closed book with partial credit to online, open book, multiple choice exams.

There is nothing here that appears negative. The response in itself is predominantly neutral but there are components that could be construed as positive and nothing that is negative. This imbalance pushes the compound score positive. The follow-up response in the Fall of 2020 had a score of −0.9023:

> I taught several introductory physics mechanics labs online in the fall. The biggest problem I found was that several students obviously doctored their results to match expectations. I don't know a good way to stop such cheating or what to do about it when I suspect it.

A very good example of a negative sentiment, the combination of words we suspect to cause negative shifts in sentiment such as "cheating" or "doctored" are within a response that contains a phrase, "biggest problem…", which we view as skewing much of the following words negative.

D. Thematic analysis

To address RQ2, the thematic analysis extracted four topics from both the initial and follow-up responses. The initial themes, as seen in Table III, were as follows: course

TABLE III. Themes in Spring 2020 and Fall 2020.

Term	Major themes
Spring 2020	Course modifications Negative aspects of the transition—primarily with labs and cheating Exam and evaluation difficulties Difficulties with student understanding
Fall 2020	Differences between Fall and Spring Cooperative learning strategies Strategies that worked in the remote space Benefits of in-person labs

modifications and technology, negative aspects of the transition—primarily with labs and cheating, Exam and evaluation difficulties and difficulties with student understanding. The extracted themes in the follow-up survey were as follows: differences between Fall and Spring, cooperative learning strategies, strategies that worked in the remote space, and benefits of in-person labs. Table IV gives example responses associated with each theme.

One aspect of LDA that we can use is a common words function for individual topics or themes. This outputs words that the unsupervised LDA algorithm has grouped into themes or topics; top words for each topic are listed in Table II. It is interesting to note that words repeat in different topics; while this may immediately raise concerns, it should be expected that words that are frequently used can occur in multiple or all of the topics. With that said, we do want to see that the words are not in the same order of dominance and that other words are interspersed. Indeed, a reason for removing words in the preprocessing phase is to eliminate words that would likely be the top word in every topic.

Some expected and unexpected results come about when comparing the themes, sample responses dominated by those themes, and the compound sentiment score from those responses. For example, the sample responses for the Spring 2020 themes of course modifications, negative aspects of the transition, and exam and evaluation difficulties all had compound sentiment scores that seem to line up with expectations. Course modifications in particular is an interesting theme because it does not have as distinct of a polarity as, for example, exam and evaluation difficulties. It is reasonable to predict both positive and negative compound scores could result from this theme. Indeed, we see a fairly positive example in Table IV. The final theme from Spring 2020, difficulties with student understanding, had a slightly positive compound sentiment, even though we

TABLE IV. Response snapshots for each theme, with selected relevant text within a response dominated by that theme and sentiment score for the included example response.

Term	Theme	Example response	Compound sentiment
Spring 2020	Course modifications	Zoom proved useful. I was able, using chat rooms, to divide students in groups and basically follow our interactive, studio-style approach with little modification. Students were responsive and were willing to participate and talk to each other. Tests were modified to reduce duration…	0.753
	Negative aspects of the transition	The most disappointing aspect is the students' willingness to cheat on exams using access to groups like Chegg, OpenStax and Google… In this small course, 75% of the students in the course joined a GroupMe where they shared the same "wrong" answers… The shared results were easy to track in the small class and those students were given an F for the course and put on Academic Probation…	−0.7495
	Exam and evaluation difficulties	Exams became open-book online, because we did not have remote proctoring software initially available and did not want honest students disadvantaged. This reduced the incentive to understand principles. Instead I saw an even greater reliance on trying to "pattern match" exam questions with previous problems. Hopefully we will have a reliable remote proctoring system in place for the fall and can return to the "single formula sheet" rule for exams.	−0.1591
	Difficulties with student understanding	I think that 5%–10% of students at my university could do well with online learning. The rest do not have the self discipline to learn physics by online courses. My fellow faculty and I saw class attendance of online lectures drop by 75% after transition to online learning. Also, about 35%–50% of students simply stopped turning in assignments or taking tests. It seems that without someone to hold their hands in-person through the process of learning, the majority of students at my school flounder…	0.4975

(Table continued)

TABLE IV. *(Continued)*

Term	Theme	Example response	Compound sentiment
Fall 2020	Differences from Spring to Fall	The biggest difference was that in Spring 2020… I taught all students face to face before and knew them and they knew me… The rest of the spring semester when we went online there was a sense of community and students were actually happy to be together, at least virtual. We all worked together to make the best of the situation. I found it much harder to connect with the students I only knew from online instruction. Also the motivation of the students was lower in Fall 2020 and many seemed to have the intention of just “checking the course off” and having it done, rather then intending to actually learn…	0.9423
	Strategies that worked in the remote space	Having had a chance to observe the technological issues the students (and we!) had during the transition to remote learning in Spring 2020 and with more time to prepare resources and strategies to help, we basically became better at helping the students manage technological issues associated with remote learning… Because our remote classes were synchronous, we were able to maintain the general approach we used during normal in-person classes.	0.9694
	Cooperative learning strategies	To make Zoom breakout groups work for cooperative learning (e.g., pair share), you need to provide clear “ice breaker” questions to get them talking. And you need to monitor continually.	0.3818
	Benefits of in person labs	In Spring 2020, I taught an introductory physics class, so we went 100% online. In Fall 2020, I was teaching our undergraduate advanced lab class. The university allowed limited in-person instruction for Fall 2020 in cases like lab and studio classes, so our department head, chief lab instructor, and I developed a plan to maintain the critical in-person hands-on laboratory part of the class…	0.915

expect this category to be dominated by negative compound sentiments. This is very interesting and we are not entirely sure of the reason for this discrepancy.

VI. DISCUSSION

Our analyses of the sentiment scores revealed that faculty who responded to the prompt were measured to have slightly positive sentiment in their written responses in our initial survey. These positive sentiments carried through to the follow-up survey, with faculty showing no shifts in sentiment. Despite the lack of shifts in sentiment, the thematic analysis did indicate shifts in what faculty saw as important in teaching remotely.

Our results support a conclusion to RQ1 that there was no statistically significant difference in the average sentiment from the initial to follow-up survey. This lack of change was surprising because instructional change is typically challenging for faculty. However, there are a large and difficult to determine number of variables that affect an individual’s experience. Further, we were surprised to see that the sentiment scores skewed somewhat more positive in both the initial and follow-up surveys.

One possible explanation for these data is that there is inherent bias in sentiment scores due to limiting the used responses to individuals who responded to both the initial and follow-up surveys. A potential reason for this bias is that faculty who responded to the survey suffered fewer consequences from the pandemic than the nonresponders and therefore had stronger feelings in one way or another. Another area of bias could come from instructors who felt distinctly different from the initial survey feeling less willing to relive or respond in a follow-up. Another potential source of bias could come from the question itself, the question asked was entirely a prompt to allow subjects to say whatever they wished. There was almost no direction prompting the subjects to respond about a particular subject. Further interesting effects could be caused by some incredibly short responses. For example, there were responses that were so short they were unable to be picked up by thematic analysis. Two examples of these would be a response that simply had the one word “horrible,” while another response was more aimed at the researchers themselves: “thank you for doing this.” These responses would show up on the sentiment analysis in very different ways while not necessarily showing up strongly in the thematic analysis.

There were nonetheless many instructors who had dramatically positive or negative sentiments on either, or both, of their responses. There are many possible reasons for these extreme sentiment scores including external impacts on the instructors like how their individual states, cities, or institutions dealt with the pandemic. Personal experiences could have a major impact as well with different people being affected dramatically different by the pandemic.

The analysis was limited to 100 data points, which is smaller than typical NLP studies. This small size limits our ability to see if there are any internal trends in the data. For example, with a larger dataset, we could separate the scores from different states and analyze on a state-to-state basis. Because states had so many different factors playing into their COVID responses, location might substantially affect faculty's experiences.

A. Sentiment programs

One of the most significant limitations of the methodology is that many sentiment analysis tools lack transparency; even open source programs, like VADER, are internally complex. VADER was found to be very widespread and accurate to general trends. With these sentiment analysis natural language processing programs, it can be difficult to identify and adjust the exact metrics being used, such as how negation and phrasing affect the sentiment. Prior to matched pair analysis, the data were run through a separate sentiment analyzer (Textblob) [24] with similar results, allowing us to move forward with VADER confidently. With that said, we do want to work with more sentiment programs in an effort to solidify the results and create a result that is independent of sentiment analysis program.

There appears to be an inconsistency in results between the sentiment analysis and the thematic analysis. Sentiment analysis shows that the median sentiment for Spring was neutral to slightly positive. However, when looking at the extracted themes, this would appear to be inconsistent with the Spring 2020 themes. As far as the consensus on topics was concerned, these themes are more negative than we expected given the sentiment analysis results. This somewhat parallels the results of Gavrin [14], which included sentiment analysis of student comments about the transition to online teaching. That sentiment analysis did not find a preponderance of negative emotion words, but hand coding the comments scored them consistently negative.

We speculate as to why the results from sentiment analysis are divergent from the LDA analysis. There are methodological differences in the two approaches that immediately draw attention, particularly in the areas of input data segmentation, word removal, and the amount of human input. Sentiment analysis is run on the raw text data, while thematic analysis is run on cleaned and abridged data. The input of LDA is a bag of words matrix, which splits every response into single words and separates them into their own columns. Sentiment analysis, however, does not simply sum up the individual sentiment of each word in a given response. There is nuance within sentiment analysis to identify commonalities in language such as negation, and how specific words affect the others around them. This was seen in the previously described did not enjoy example.

Thematic analysis assumes that many of the removed structures such as punctuation and stop-words are not useful in the determination of the topic vectors [19]. Sentiment analysis is affected by the inclusion of these removed phrases and punctuation. We did remove words from the responses when applying thematic analysis algorithms in order to more properly identify the latent themes; these included “class,” “online,” “teach,” and “time,” These words are more likely to have an effect in the sentiment analysis by inclusion in n-grams than stop words.

The LDA analysis, due to the smaller dataset, required more human coding input and topic determination. There was a less clear distinction from a computation aspect as to whether a three or four topics were the best choice when it came to the Fall responses, for instance. This required us to interpret the data and provide a solution manually. The sentiment program, however, did not require any human intervention in its operation outside of data processing and cleaning. With the parallels mentioned in the results of Gavrin [14], it is possible this human intervention on LDA analysis, with choosing topic number and then the group effort to make topic determination, that there exists an explanation for the discrepancy in our results.

It remains unclear what weight each of these differences in approaches has on the overall analysis of the text data, and we are not confident in labeling one more “accurate” or “correct” than the other. With all machine learning text-based analysis, we return to the fourth principle by Grimmer and Stewart “There is no globally best method for automated text analysis” [12]. Our results demonstrate the usefulness of contrasting multiple text processing methods, because each may highlight different aspects of the data.

B. Themes

The remaining two research questions deal with what themes emerged from survey responses (RQ2) and how they changed from Spring to Fall (RQ3). In Spring, the themes suggest that instructors were dealing with the difficult task of dealing with an online environment that they had little, or no, experience with. Initially, this suggests that instructors struggled with the aspects of instruction that are the most well developed and the most traditional. We identified themes relating to how to properly evaluate individuals when you could no longer have them all in a single room working on a piece of paper and could control the access to information available during exams. The majority of these traditional exam systems are not available when dealing with online instruction. While there is a large amount of online instruction that existed before the pandemic, evaluation resources have lagged. This is particularly true for disciplines that have resisted online instruction in the past, like within the STEM fields.

We saw that cheating was a large concern for instructors; the lack of control that instructors had over what their online students are able to do during examination proved to be a large concern. This is shown in the themes relating to cheating and exam difficulties. While these two themes appear to be the same, we see a distinction between the negative aspects of the transition and exam difficulties. For example, cheating appeared within both themes, however, the exam difficulties brought up cheating more frequently in regard to how technology could be used or failed to detect cheating. Meanwhile, the negative aspects theme showed cheating as something the instructor had distinct experiences with, typically within their class, and how that influenced their experiences.

Along with cheating, instructors displayed negative emotions about labs. Labs are often well thought out and institutions often spend large amount of time and effort to develop labs for their courses. Having to remake labs to work online is a difficult proposition given ample time and resources. In the case of many institutions, the switch to online lab instruction was made with a few weeks notice and they had little to no remote lab resources already developed.

The final theme in the initial data was instructors expressing their concern regarding information retention in students. Anecdotal stories within the responses led instructors to believe that their students did not retain the material as well as previous courses, and that the technology was not conducive to the appropriate learning environment.

Fall showed a different story when the themes are extracted. We expected the first theme, comparisons between the previous Spring semester and the current Fall semester. The second theme, cooperative learning strategies, demonstrates the benefit of having Spring to learn from. This theme focused on the benefits and ways that group work can be incorporated into the course. Common examples involved both technological strategies such as Zoom breakout rooms as well as group projects worked into courses to replace standard exams.

The third topic revolved around the strategies that instructors have been using that are effective. It is likely that having more time and familiarity with virtual learning allows instructors to develop and implement successful learning strategies. The last topic of the follow-up surveys was the only one that had similarity to a topic in the initial survey, which being the benefits of in-person labs. During the Fall of 2020, there were some institutions that began to bring students back into campus, particularly undergraduates and lab-based science classes. The Fall lab-based theme focused primarily on how difficult and ineffective it is to conduct physics labs in the online environment. The Spring lab-based theme echoes this theme by focusing on how being back in person for labs has a dramatic positive effect on the lab work and student engagement. The way that labs are discussed in this section indicates that they were the most difficult type of instruction to effectively navigate in the online environment.

C. Closing remarks

We hope that the two methods outlined in this paper, sentiment analysis and thematic analysis, can open up new avenues for investigations within physics education research. These methods offer a novel way of approaching the analysis of qualitative data, particularly large datasets where manual evaluation is human-cost heavy. The differences between our sentiment and thematic analysis results also give an important caution to cross-check between different techniques and include manual inspection of the data. By using the large-scale processing abilities of computers to highlight trends of interest, we can better target the resources of human researchers even in large datasets. It is not yet clear what the “new normal” of physics education will become in the post-COVID world, but there is a great need for research tools that can adapt to 21st-century data. A possible interpretation of these results is to try to identify one methodology as being more sound or “correct” than the other. This can be a tricky avenue of discussion because each methodology has pros and cons that could be weighed differently to different researchers. For example, the amount of human intervention required for our LDA analysis could make it more attractive to certain researchers; however, others may view the sentiment analysis, with its enormous pretrained database and wide array of use to be the more attractive methodology. This is something that we did not set out to facilitate, the goal of this research is not to push one methodology over the other but rather to highlight the different branches of machine learning (supervised and unsupervised) and their possible uses in PER. We present these results with the explicit connection to their machine learning methodologies, the results should be intrinsically connected to, and discussed with, the methods applied to achieve them. As we write this paper, the world of machine learning and natural language processing is exploding as the availability and functionality of these algorithms become increasingly favorable. Modern tools such as Weka [35] make the computational side of these large models accessible for a wider range of researchers. Within NLP alone, there are advances in combining different methods and models, for example, experimentation of putting word embeddings into LDA [36]. These changes demonstrate the push for these machine learning tools to play a more active role in qualitative and, like ours, quantitative research.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation Awards No. DUE 2027958 and No. DUE 2027963. We appreciate the methodological help from Tor Ole Bigton Odden on Latent Dirichlet Allocation. We also greatly appreciate the time all survey respondents spent, particularly during a pandemic.

[1] K. Lewin, Frontiers in group dynamics: Concept, method and reality in social science; social equilibria and social change, Hum. Relat. **1**, 5 (2016).

[2] A. Werth, C. G. West, and H. Lewandowski, Impacts on student learning, confidence, and affect and in a remote, large-enrollment, course-based undergraduate research experience in physics, Phys. Rev. Phys. Educ. Res. **18**, 010129 (2022).

[3] E. Brewe, A. Traxler, and S. Scanlin, Transitioning to online instruction: Strong ties and anxiety, Phys. Rev. Phys. Educ. Res. **17**, 023103 (2021).

[4] J. L. Gordon and J. Presseau, Effects of parenthood and gender on well-being and work productivity among Canadian academic research faculty amidst the COVID-19 pandemic, Can. Psychol./Psychologie canadienne, **64**, 144 (2023).

[5] P. Klein, L. Ivanjek, M. N. Dahlkemper, K. Jeličić, M.-A. Geyer, S. Küchemann, and A. Susac, Studying physics during the COVID-19 pandemic: Student assessments of learning achievement, perceived effectiveness of online recitations, and online laboratories, Phys. Rev. Phys. Educ. Res. **17**, 010117 (2021).

[6] L. S. Neuwirth, S. Jović, and B. R. Mukherji, Reimagining higher education during and post-COVID-19: Challenges and opportunities, J. Adult Contin. Educ. **27**, 141 (2021).

[7] L. Lepp, T. Aaviku, Ä. Leijen, M. Pedaste, and K. Saks, Teaching during COVID-19: The decisions made in teaching, Educ. Sci. **11**, 47 (2021).

[8] K. Chowdhary, *Fundamentals of Artificial Intelligence* (Springer, New York, 2020), pp. 603–649.

[9] J. Hirschberg and C. D. Manning, Advances in natural language processing, Science **349**, 261 (2015).

[10] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, Natural language processing of social media as screening for suicide risk, Biomed. Inf. Insights **10**, 1 (2018).

[11] D. Hotiana, B. Duncan, D. Tamma, and W. Courtney, NLP approach for mental health problems associated with social media activities (2021), http://hdl.handle.net/1920/12190.

[12] J. Grimmer and B. M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts, Political Anal. **21**, 267 (2013).

[13] P. Kelley, A. Gavrin, and R. S. Lindell, Text mining online discussions in an introductory physics course, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2017.pr.049.

[14] A. Gavrin, Physics students' reactions to an abrupt shift in instruction during the COVID-19 pandemic, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.Gavrin.

[15] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing—Volume 10* (Association for Computational Linguistics, USA, 2002), pp. 79–86.

[16] L. Kurniasari and A. Setyanto, Sentiment analysis using recurrent neural network, J. Phys. Conf. Ser. **1471**, 012018 (2020).

[17] H. Newman and D. Joyner, Sentiment analysis of student evaluations of teaching, in *International Conference on Artificial Intelligence in Education* (Springer, New York, 2018), pp. 246–250.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. **3**, 993 (2003).

[19] T. O. B. Odden, A. Marin, and M. D. Caballero, Thematic analysis of 18 years of physics education research conference proceedings using natural language processing, Phys. Rev. Phys. Educ. Res. **16**, 010142 (2020).

[20] AIP Statistical Research Center, https://www.aip.org/statistics.

[21] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media, Inc., 2009).

[22] C. Hutto and E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in *Proceedings of the International AAAI Conference on Web and Social Media* (2014), Vol. 8, pp. 216–225, 10.1609/icwsm.v8i1.14550.

[23] V. Bonta and N. K. N. Janardhan, A comprehensive study on lexicon based approaches for sentiment analysis, Asian J. Comput. Sci. Technol. **8**, 1 (2019).

[24] S. Loria, Textblob documentation, Release 0.15, Vol. 2, 2018.

[25] D. Navarro, Learning statistics with R: A tutorial for psychology students and other beginners: Version 0.5, University of Adelaide Adelaide, Australia, 2013.

[26] R. D. Morey, J. N. Rouder, and T. Jamil, Bayesfactor: Computation of Bayes factors for common designs. R package version 0.9. 12-4.2 (2018).

[27] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021, https://www.R-project.org/.

[28] R. Rehurek and P. Sojka, Gensim–python framework for vector space modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3**, 2 (2011).

[29] S. Kapadia, Evalute topic models: Latent dirichlet allocation (LDA), 2019, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

[30] S. Syed and M. Spruit, Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation, in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, New York, 2017), pp. 165–174.

[31] The pandas development team, pandas-dev/pandas: Pandas, 2020, 10.5281/zenodo.3509134.

[32] W. McKinney, Data structures for statistical computing in python, in *Proceedings of the 9th Python in Science Conference* (Austin, TX, 2010), Vol. 445, pp. 56–61, 10.25080/Majora-92bf1922-00a.

[33] C. May, R. Cotterell, and B. Van Durme, An analysis of lemmatization on topic models of morphologically rich language, arXiv:1608.03995.

[34] R. E. Kass and A. E. Raftery, Bayes factors, J. Am. Stat. Assoc. **90**, 773 (1995).

[35] M. A. H. Eibe Frank and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques," Fourth Edition* (Morgan Kaufmann, 2016).

[36] F. Esposito, A. Corazza, F. Cutugno *et al.*, Topic modelling with word embeddings, CLiC-it/EVALITA (2016).