Leveraging Institutional Data to Understand Student Perceptions of Teaching in Large Engineering Classes

Michelle Soledad
Department of Engineering Education / Electrical Engineering Department
Virginia Tech / Ateneo de Davao University

Jacob Grohs and Sreyoshi Bhaduri
Department of Engineering Education
Jennifer Doggett
Department of Mining & Minerals Engineering
Jaime Williams and Steven Culver
Office of Assessment and Evaluation
Virginia Tech

***Abstract*—A global push to pursue careers in engineering has led to an increase in enrollment in engineering programs. However, rising student populations have led institutions to make compromises in order to effectively manage existing resources and rising costs, such as resorting to large classes despite evidence that they may be detrimental to student learning. Recognizing that large classes are both a necessity for institutions and a challenge for the instructors who teach them, we seek ways to help faculty create effective learning environments despite the difficulties posed by this setting. Developing an effective learning environment requires instructors to reflect and consider input from various sources, including students. A source of data for student input are student perceptions of teaching surveys. This paper used the MUSIC Model of Academic Motivation as basis to characterize qualitative data from student surveys with respect to two of the five MUSIC dimensions: *Success* and *Caring.* We allowed categorical variables to emerge from qualitative data and investigated how quantitative results from the student evaluation (e.g., *Did the instructor present the material clearly?)* varied across categories. The manually-analyzed text data were also used to explore text analytics as a qualitative analysis technique for course evaluation surveys.**

Keywords—large classes; student surveys; student evaluation; MUSIC Model; Natural Language Processing; bag-of-words model

I. INTRODUCTION

A global push to pursue careers in engineering has led to an increase in enrollment in undergraduate engineering programs [1]. While this in itself is a positive development, rising student populations have led institutions to make compromises in order to effectively manage existing resources and rising costs [2]. An example of such a compromise is resorting to large class sizes despite evidence that they may be detrimental to student learning [3].

Fundamental engineering courses provide the foundation upon which advanced discipline-specific courses are built [4]. These courses are often taken concurrently and required of multiple majors (e.g. [5]). Statics, for example, is required in the Aerospace, Biological Systems, Biomedical, Civil, Construction Management, Environmental, Industrial Operations, Industrial Systems, Materials Science, Mechanical, Mining and Ocean Engineering programs, as well as in Engineering Physics and Engineering Science & Mechanics [6]–[10].

Fundamental courses (such as Statics) that are required of multiple disciplines during a stage in the college journey that is marked by collectively high enrollment rates are more likely to be organized and taught in large class sizes [2], [11], [12] because they provide an opportunity to maximize faculty contact hours and institutional resources. Large classes, however, have been associated with unfavorable educational environments because of decreased meaningful interaction between instructors and students, among other reasons [3]. Seymour and Hewitt [12] also identified the occurrence of large class sizes as one of the reasons why students choose to leave engineering, although it is not as strong a predictor of the decision to leave as the perception that a non-engineering major would be able to provide a "better educational experience." Instructors, on the other hand, have expressed that it was challenging to create a positive learning experience for students in a large class [13]–[15] and ensure that students have access to individual help [16].

Recognizing that large classes are both a necessity for institutions and a challenge for the instructors who teach them [13], [14], we seek ways to help faculty create effective learning environments despite the difficulties posed by this setting. Developing an effective learning environment requires reflection on the part of the faculty and input from various sources, including students [17]. A source of data to inform this process are student evaluations or student perceptions of teaching surveys.

Student evaluations are administered regularly by institutions for non-research purposes (e.g., evidence for the tenure and promotion process). It may often be the only formal avenue for students to provide feedback regarding their learning experience. In theory, student evaluations contain information that should help faculty acknowledge and respond to the student voice when they make curricular and pedagogical decisions. Student evaluations, however, have been deemed underutilized and unreliable in some literature [18]. Analysis is usually limited to calculations of numeric ratings, and responses to open-ended items are presented as a list of all the comments collected from the respondents, copied verbatim. Despite concerns over

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

reliability, there are scholars who argue that the limitations of course evaluation surveys may be addressed by ensuring that survey instruments undergo the appropriate level of psychometric testing [19]. We also recognized the opportunity to leverage readily-available data (in the form of text responses to open-ended survey items) that is otherwise not being used for research purposes, and is currently not being analyzed, in a more meaningful way.

II. Student Evaluations and the MUSIC Model

A. Student perceptions of teaching surveys: reliability and potential significance

Institutions gather data on student experiences through course evaluation and student perceptions of teaching (SPOT) surveys. Students are given an opportunity to anonymously describe and evaluate their learning experience in the course. These observations are then shared with the instructor and may ideally be used to inform curricular and pedagogical decisions. Unfortunately, this potential is usually not maximized [18]. We find value, however, in the opportunity to explore how to synthesize and present text responses to open-ended SPOT survey items in a form that is more constructive and useful to the instructors of large fundamental engineering classes.

Criticisms of SPOT surveys revolve around insufficiency, inability to take into account the multidimensional nature of teaching, and gender bias [19]–[21]. Some engineering faculty lamented that the results of these surveys may not be accurate indicators of teaching quality, and may be used inappropriately against them [22]. Marsh and Roche (1997) argue, however, that student evaluation instruments that "fail to provide a comprehensive evaluation of theoretically sound, multiple dimensions of teaching quality" are those that have not undergone the appropriate level of evaluation relative to rigorous psychometric considerations [19]. It is, therefore, possible for SPOT surveys to be multidimensional, reliable, stable, valid, and useful in improving teaching effectiveness, as long as the instrument is developed with appropriate consultation and subjected to rigorous development practices.

We believe that all these concerns are valid, especially in the absence of transparent information on how instruments were developed and if interpretation of the survey results is limited solely to the quantitative information derived from the data (e.g., student-provided ratings). This study, however, did not use SPOT evaluations as a measure of teaching quality or effectiveness; the intention was not to evaluate the instructor or the course. Instead, SPOT responses were used as a window into how students articulated their learning experience and what their perceptions and interpretations of their instructor's attitudes and behaviors were. Qualitatively analyzing open-ended responses provided an additional layer of processing of data, and ascribed deeper meanings and interpretations of student perceptions. Our goal is to contribute to the more constructive use of SPOT evaluations so that these instruments can truly contribute to the improvement of the student learning experience.

B. Qualitative Analysis of SPOT and the MUSIC Model of Academic Motivation

In a previous study, we qualitatively analyzed student responses to open-ended items in SPOT surveys of Mechanics courses in a large public research institution [23]. Our analysis identified the following constructs: *facilitator of learning, quality of interaction, self-regulated learning,* and *physical environment.* We also found that the emergent codes and themes related to the first three constructs *(facilitator of learning, quality of interaction,* and *self-regulated learning)* indicated that students preferred learning strategies suggested by the MUSIC Model of Academic Motivation.

The MUSIC Model of Academic Motivation was developed by Jones [24]. It is the outcome of analyzing, evaluating, and synthesizing research and theories on academic motivation. The MUSIC Model consists of five components: *eMpowerment*, *usefulness*, *success*, *interest* and *caring.* Dr. Jones envisioned these components as what *"an instructor should consider when designing instruction"* [24] in order to keep students motivated and engaged in the learning process.

The codes generated from our previous study [23] could be clustered according to the MUSIC Model components. They also aligned with strategies designed to motivate students by satisfying student need for the dimensions of the MUSIC model [25]. These findings indicate the potential use of the MUSIC Model of Academic Motivation [24] as a qualitative framework to present open-ended responses to student surveys in a meaningful manner and provide instructors with strategies to motivate students, promote student engagement, and improve learning.

C. Text Analytics as a Qualitiative Analysis Technique

As previously stated, our objective is to analyze the responses to open-ended items in SPOT surveys as a window into how students articulate their learning experiences and what their perceptions and interpretations of their instructors' attitudes and behaviors are. This analysis may be particularly useful to instructors since it gives them a perspective into their students' learning experiences in their course and allows them to reflect on the quality of their interactions with students from the students' point of view. At this time, the dissemination of SPOT survey results to instructors consist of presenting means and frequencies for the numeric responses and a list of the responses to open-ended responses, copied verbatim. For instructors of large classes, the information from open-ended survey items translates to a relatively large volume of text-based responses that range from single-word statements to paragraph-long rants. Going through these responses is time and resource-intensive for individual instructors to manually analyze qualitatively.

Text analytics through use of Natural Language Processing may be useful in such situations. Automated textual analysis can be successful in extracting meaningful data from large volumes of text, and thus help inform research, practice and pedagogy. Natural Language Processing (NLP) techniques may be used in conjunction with machine learning tools to enable automation of analytical processes such as tasks related to automatically classifying texts or automatically identifying primary sentiment conveyed. Natural Language Processing can

thus be understood as an interdisciplinary field in which computers are used to perform useful tasks involving human language [26].

NLP techniques have been used in conjunction with machine learning tools for varied content analysis related tasks. One of the earlier definitions of machine learning was provided by Samuel [27] who stated that these types of studies are concerned with the "programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning" (p. 71) in describing machine learning using a game of checkers.

Examples of text-based content analysis include the use of supervised machine learning techniques of classification with Natural Language Processing techniques to categorize suicide notes as genuine or illicit [28] and to analyze over 300 State of the Union addresses by Presidents of the United States of America to determine a timeline of trends in topics of national interest [29]. In the instance of suicide note classification described by Pestian, et al [28], it was determined that classification algorithms outperformed manual classification by mental health professionals. The algorithm correctly classified whether a suicide note was genuine or not 78% of the time as opposed to 63% when classification was done manually by mental health professionals.

In this study we employ the NLP technique of bag-of-words modelling and a machine learning based classifier to demonstrate how analysis of open-ended text-based data of SPOT surveys can be automated to mimic researcher's qualitative coding.

III. Data and Methods

This study is part of an ongoing effort, consisting of individual studies, to understand the learning experience in large classes by leveraging institutional data. In a previous study, we analyzed text data from SPOT surveys and recognized the following potential future work: 1) the use of the MUSIC Model as a framework for clustering and presenting responses to the open-ended items informatively and constructively, and 2) the development of an automated technique for processing text responses to SPOT surveys [23], [30]. This study pursued these ideas by qualitatively analyzing SPOT responses from Mechanics courses using the MUSIC Model of Academic Motivation as a coding framework, then using the manually-coded data to explore automated qualitative analysis through natural language processing techniques. We limited the scope of work to a portion of a purposefully-selected sample of SPOT responses and the two most prevalent MUSIC components from the previous study: *Success* and *Caring*. We chose this approach in order to see how viable the results are and to use these results to design a more refined and robust system for processing text data.

A. Description of Data

Our data consisted of responses to SPOT surveys for all offerings of Statics, Dynamics, and Strength of Materials in a large public research university over a period of four consecutive semesters (Fall 2014, Spring 2015, Fall 2015 and Spring 2016). The data (3,917 responses) was provided to us by the office that administered student surveys. The records did not contain information that identified the student-respondents; in cases where answers identified course instructors by name, pseudonyms were generated to ensure the instructors' anonymity. Our protocol for preparing and analyzing existing data was subjected to review and approval by our Institutional Review Board.

SPOT surveys periodically collect data for non-research purposes (e.g. generating feedback for instructors, evidence for tenure and promotion) "but contain rich, descriptive information" [30] that may be analyzed for research purposes. Using existing data such as student survey responses for original research is a way to maximize the usefulness of previously-collected data [31].

We employed purposive sampling [32] to select responses for analysis, in order to ensure that we are able to qualitatively analyze answers to open-ended survey items across all courses and semesters. The sampling process chose responses from each instructor, course offering, and semester that contained an answer to at least one of the following open-ended survey items:

What did the instructor do that helped the most in your learning? (Q7)

Please add any additional comments regarding the course and/or instructor here. (Q9)

The sampling process resulted in the selection of 1,262 responses that met our purposive sampling criteria, guided by content analysis methodology. Content analysis is a method used to analyze text data from open-ended survey items and generates a "subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes and patterns" [33].

B. Qualitative Content Analysis of Answers to Open-Ended Items

We analyzed data from the purposefully-selected sample that consisted of responses from the Fall 2014 semester (*325 responses*) only. We used existing codes for the *Success* and *Caring* MUSIC dimensions identified in our previous study [23] and the strategies for implementing the MUSIC Model in the classroom [24] to qualitatively analyze answers to the open-ended survey items listed in Section III.A above.

The answers to open-ended survey items for the Fall 2014 semester were qualitatively analyzed manually by two investigators independently and concurrently. Analysis focused only on the *Success* and *Caring* components and exclusively followed *a priori* codes for these constructs. The codes were derived from the suggestions offered by the MUSIC Model of Academic Motivation [24], [25] when designing for *Success* and *Caring* in the classroom (15 codes for *Success*, 7 codes for *Caring*). The coding process considered and labeled answers that embodied both positive and needs improvement instances of the constructs. Memos and notes were also generated during analysis. After the independent coding process, the investigators met, discussed, and consolidated the results of the qualitative analysis. Positive and needs improvement instances of *Success*

and *Caring* were clustered into buckets and quantified using frequency counts.

The manually-analyzed dataset from the Fall 2014 semester was then used to train the classifier algorithm. The first step in the Natural Language Processing was *pre-processing,* which included preparing the data for analysis. At first, all of the punctuation marks are removed from the document and all the words are converted to lower case. After that, each of the words were lemmatized using Word Net [34], [35]. The responses to questions on the SPOT survey, along with their associated labels for *Success* or *Caring* that were assigned during the manual coding process, was used to train a bag-of-words model to automate the classification process. Our end goal was to develop an algorithm with a high accuracy of classifying a segment of response as either indicative of *Success* or *Caring*.

C. Analysis of Selected Likert-type Survey Items

Descriptive statistics were generated for the quantitative items of responses that contained answers to open-ended items coded as *Success* and *Caring*. Analysis anchored on *Success* and *Caring* was done on the following items that required students to rate the instructor and/or course using a 6-point Likert scale:

TABLE I. LIKERT-TYPE SURVEY ITEMS INCLUDED IN ANALYSIS

Survey Item	MUSIC Dimension
The instructor presented the subject matter clearly	Success
The instructor provided feedback intended to improve my course performance	
I have a deeper understanding of the subject matter as a result of this course	
I improved my ability to problem solve	
The textbook or course readings made a valuable contribution to my learning	
The objectives of the course were clearly explained	
The instructor fostered an atmosphere of mutual respect	Caring

Means for the following item were also generated to indicate a respondent's general evaluation of the course instructor: *Overall, the instructor's teaching was effective.*

D. Categorical Classifications and Variation Across Categories

The results of qualitative analysis of open-ended items and quantitative analysis of Likert-type items were categorized using the following criteria:

TABLE II. CATEGORICAL CLASSIFICATIONS FOR *SUCCESS* AND *CARING*

Category	Type of Survey Item	
	Open-ended	***Likert-type***
High	Positive instances are 75% or higher of total coded answers	Mean of 4.6 or higher
Medium	Positive instances are below 75% but above 35% of coded answers	Mean lower than 4.6 and higher than 2.6
Low	Positive instances are 35% or lower of total coded answers	Mean of 2.6 or lower

The criteria categorized survey responses as *high, medium* or *low* on *Success* and *Caring.* These classifications were then used to investigate how Likert-type responses varied across categories, comparing against answers to open-ended items.

IV. RESULTS

This study sought to: 1) qualitatively analyze answers to open-ended items based on the *Success* and *Caring* dimensions of the MUSIC Model; 2) investigate how quantitative answers to Likert-type survey items varied across *Success* and *Caring* categories of the qualitative data; and 3) explore automated qualitative analysis of text responses to SPOT surveys using natural language processing techniques. For the manually-analyzed Fall 2014 semester data, a total of 808 statements across two open-ended items were labelled with codes categorized as *Success* and *Caring*.

Table III below shows the distribution of coded statements. These statements were coded and identified from 325 individual responses from the Fall 2014 semester sample. The two investigators examined each response, identified specific keywords or phrases as perceptions related to *Success* and *Caring*, and classified them as either positive or needs improvement instances of the construct. One response may include multiple coded statements, and each statement may be coded simultaneously. *Simultaneous coding* refers to assigning more than one code to a phrase or statement [36]. The possibility of having multiple coded statements per response and having multiple codes for one statement led to a higher number of coded statements (808) vis-à-vis the number of responses that were analyzed (325).

TABLE III. STATEMENTS CATEGORIZED AS *SUCCESS* AND *CARING* FALL 2014 SEMESTER

Success		Caring		Total
Positive	***Needs Improvement***	***Positive***	***Needs Improvement***	
444	184	160	20	808
Total Coded Statements, *Success*				628
Total Coded Statements, *Caring*				180
% Positive Instances, *Success*				70%
% Positive Instances, *Caring*				86%

Most of the responses talked about perceptions and experiences related to *Success* (628). For the most part, the text responses recounted positive comments and experiences for both *Success* (70%) and *Caring* (86%). The percentage of positive statements were calculated and used as basis for categorizing coded statements as *high, medium,* or *low* for *Success* and *Caring*,

A. Results of Qualitative Analysis

Our qualitative analysis used suggestions to design for *Success* and *Caring* in the classroom [24] as the coding frame. Responses were analyzed and coded as *Success* and *Caring* based on whether student perceptions were associated with a suggestion related to that construct. Each coded statement was then identified as a *positive* or *needs improvement* instance for

the construct it describes (*Success, Caring*). A response yielded one or more coded statements, and may simultaneously be labelled with two or more codes.

1) Success: Statements coded under this dimension talked about student perceptions of how faculty contribute to, influence, and increase their capability to succeed in the course and build a good learning environment. Examples of positive instances for *Success* include employing cognitive modeling strategies in the classroom and student perceptions of commitment on the part of the instructor to ensure understanding of the course material: *"Professor Unice is always available to answer questions. She teaches very well, but if someone doesn't understand a concept she will stop and take the time to make sure that everyone understands."*

Survey answers indicate that students associate such faculty behavior as providing guidance on how to think about and approach work out problems and providing meaningful, timely feedback with a positive learning experience and higher probability for success in the course. These strategies align with suggestions regarding *designing for success* based on the MUSIC Model [24], [25].

Examples of needs improvement instances for *Success,* on the other hand, include perceptions of faculty behavior that seem to indicate inability to provide timely feedback and inaccessibility for questions and clarifications: *"I have never felt like Professor David did anything to help his students. He is very stoic and unresponsive to requests for help and not very good at explaining concepts. I would say that the only helpful thing he has every done is encourage me to seek help elsewhere."*

2) Caring: Statements coded under this dimension expressed student perceptions of an instructors' ability to empathize with students, and observations of instances of instructors showing compassion to students. Examples of positive instances for *Caring* included personality-based perceptions of friendliness, approachability, helpfulness, and showing concern for the welfare of students. Some statements that were characterized as *Caring* intersected, and were also coded as, statements associated with *Success*: *"Was friendly and approachable. Clearly cared about the students and allowed for student input during lectures. Worked through plenty of practice problems to help understand."* These observations were in keeping with suggestions regarding *designing for caring* based on the MUSIC Model [24], [25].

Needs improvement instances for *Caring*, on the other hand, included student perceptions of lack of respect or concern: "*He did not respect his students and did not talk to them politely*."

B. Results of Training Run for the Classifier Algorithm

The bag-of-words model is a method for text categorization tasks in which the features in the document are represented by weighted occurrence frequencies of individual words [37]. We used the model to generate the most informative features (Fig 1) to classify segments of the student responses.

As can be seen in the Figure 1, the occurrence of the word "example" indicates that a response is 15.5 times more likely to be related to the student expressing the MUSIC model construct of *Success* rather than *Caring*.

Fig. 1. *Top 20 of the most informative features to enable classification*

```
Most Informative Features
       contains(example) = True           Succes : Caring =     15.5 : 1.0
          contains(open) = True           Caring : Succes =      9.6 : 1.0
         contains(cared) = True           Caring : Succes =      7.5 : 1.0
        contains(people) = True           Caring : Succes =      7.5 : 1.0
         contains(extra) = True           Caring : Succes =      5.3 : 1.0
       contains(concern) = True           Caring : Succes =      5.3 : 1.0
       contains(william) = True           Caring : Succes =      5.3 : 1.0
      contains(possible) = True           Caring : Succes =      4.5 : 1.0
        contains(office) = True           Caring : Succes =      4.3 : 1.0
       contains(provide) = True           Succes : Caring =      4.1 : 1.0
          contains(mike) = True           Caring : Succes =      3.9 : 1.0
          contains(hour) = True           Caring : Succes =      3.9 : 1.0
      contains(practice) = True           Succes : Caring =      3.8 : 1.0
       contains(student) = True           Caring : Succes =      3.7 : 1.0
         contains(solve) = True           Succes : Caring =      3.7 : 1.0
     contains(available) = True           Caring : Succes =      3.5 : 1.0
          contains(step) = True           Succes : Caring =      3.4 : 1.0
      contains(homework) = True           Succes : Caring =      3.4 : 1.0
        contains(person) = True           Caring : Succes =      3.2 : 1.0
      contains(flexible) = True           Caring : Succes =      3.2 : 1.0
      contains(whenever) = True           Caring : Succes =      3.2 : 1.0
          contains(back) = True           Caring : Succes =      3.2 : 1.0
      contains(thorough) = True           Caring : Succes =      3.2 : 1.0
      contains(response) = True           Caring : Succes =      3.2 : 1.0
          contains(week) = True           Caring : Succes =      3.2 : 1.0
          contains(sent) = True           Caring : Succes =      3.2 : 1.0
        contains(advise) = True           Caring : Succes =      3.2 : 1.0
       contains(endorse) = True           Caring : Succes =      3.2 : 1.0
contains(answerproblem) = True           Caring : Succes =      3.2 : 1.0
            contains(mr) = True           Caring : Succes =      3.2 : 1.0
```

The bag-of-words classification model was tested on a partitioned pre-manually coded data set. The algorithm was run 100 times, each time randomly choosing 80% of the coded response to train the algorithm, and testing the algorithm on the remaining 20% of data. The average accuracy of the algorithm over 100 runs was 74.47%; this translates to the machine being able to correctly mimic the human coded labels approximately 75% of the time on an average. The highest accuracy recorded across the 100 runs was 82.8% while the lowest was 65.7%. Accuracy may have been affected by low training instances for *Caring*; since *Success* statements significantly outnumbered those categorized as *Caring*, it is possible that some statements are being falsely classified as *Success.* An average accuracy of 75%, however, may be considered *acceptable* [28]. For a first attempt using a small portion of data, this finding is promising. It means accuracy may be improved by subjecting the algorithm to more training runs using additional pre-manually coded data. We will be able to accomplish this through a succeeding study that will use data from the Spring 2015, Fall 2015, and Spring 2016 semesters.

C. Results of Quantitative Analysis

Seven survey items that required numeric responses were included in the quantitative analysis. Six of the survey items are associated with *Success* while one is associated with *Caring,* as indicated in Table I. Students are asked to indicate the degree to which they agree to the statement indicated in each item, using a 6-point Likert scale (*6=Strongly agree; 1=Strongly disagree).* Means and frequencies of the responses analyzed for each semester are shown in Table IV below.

TABLE IV. ANALYSIS OF QUANTITATIVE RESPONSES, SUCCESS & CARING

Term	Number of Respondents	Means, Success	Means, Caring	Means, Overall
Fall 2014	325	4.95	5.40	4.92
Spring 2015	324	5.07	5.52	5.25
Fall 2015	286	4.72	4.79	4.68
Spring 2016	327	4.87	5.33	4.83
	Total = 1262	4.90	5.26	4.92

The numeric responses of 1,262 respondents were included in the analysis, with approximately 300 respondents each for the Fall 2014, Spring 2015, Fall 2015 and Spring 2016 semesters. Means were calculated for *Success, Caring* and *Overall Teaching effectiveness* for each semester using numeric responses to six survey items for *Success*, one survey item for *Caring*, and one survey item for *Overall Teaching Effectiveness* (Table I). The values indicated in Table IV above represent means across respondents per semester, considering survey items included for analysis for each dimension.

Calculated means for the *Success* and *Caring* dimensions indicate that student perceptions of their instructor's classroom strategies related to these constructs are favorable. Student ratings in items related to *Success* and *Caring,* as well as *Overall Teaching Effectiveness*, are high. *Caring*, which is covered by one item (*the instructor fostered an atmosphere of mutual respect*) is rated higher than *Success.* These values, however, do not measure the extent to which the respondents perceive the presence of the *Success* and *Caring* components in their Mechanics courses as intended by the MUSIC Inventory [24]. We do not intend to measure the *Success* and *Caring* components in this manner. For this study, we use these observable variables from the SPOT survey that we have identified as related to *Success* and *Caring* to examine how latent variables that emerged from the qualitative analysis of answers to open-ended items vary across these two MUSIC components.

The following classifications emerged from the Fall 2014 semester data, based on the results shown in Tables III and IV and the criteria shown in Table II:

TABLE V. CATEGORICAL CLASSIFICATIONS

Dimension	Category (*High, Medium, Low*)
Success, Quantitiative	High
Success, Qualitative	Medium
Caring, Quantitative	High
Caring, Qualitative	High

The results shown in Table V above indicate that student perceptions related to how *Success* and *Caring* are fostered in their Mechanics courses are consistent across both quantitative and qualitative data. Student perceptions are, in general, positive, with most ratings and experiences classified as *High* on both *Success* and *Caring*. Qualitative data (statements of student experiences) on *Success*, the only construct categorized as *Medium*, was still relatively favorable, with 70% of the statements for the construct labelled as positive comments and experiences.

V. IMPLICATIONS AND FUTURE WORK

The key findings of this study were:

1) Students shared positive statements related to *Success* and *Caring* that indicate student perceptions of faculty behavior towards facilitating conceptual understanding, providing meaningful and timely feedback, cultivating an atmosphere of mutual respect, and showing concern for the quality of the students' learning experience.

These perceptions are in keeping with characteristics of an ideal engineering classroom as described in literature [38]. It is promising to find that students are able to acknowledge the effort of instructors to foster positive learning experiences. Students, however, continue to struggle in Mechanics courses [39], and bridging the gap between effort on the part of the instructor to create an effective learning environment and actual student performance in the class merits a closer look.

2) Instances of postive perceptions (604 statements) that emerged from qualitative data related to *Success* and *Caring* are associated with *High* ratings (4.9 for *Success*, 5.26 for *Caring*) for numeric survey items using a 6-point Likert scale for the same dimensions. Quantitative ratings and qualitative data from student experiences are consistent across constructs, and generally reflect favorable experiences.

Consistency across quantitative ratings and qualitative data may indicate that students who chose to participate in the SPOT surveys, at least for the institution that provided data for and the semesters included in this study, were intentional about their responses. It may indicate further that analyzing and considering responses to the open-ended items will provide additional and more specific information that will help instructors understand what the numeric ratings mean. This information may also be used to inform decisions related to choosing learning activities and designing assessment tools, among other considerations, when preparing for succeeding offerings of a course. Such ideas will not be readily identifiable from student ratings in survey items that require numeric responses.

3) A trial run employing natural language processing techniques to qualitatively analyze text responses to course evaluation surveys yielded an *acceptable* accuracy rate of approximately 75%. For an initial run, this level of accuracy is promising. Accuracy may be improved by conducting more runs to train the algorithm and expanding baseline data by adding more pre-manually coded statements. At this time, there are data that will allow us to accomplish the aforementioned

steps to improve accuracy, indicating that further development and refinements to the algorithm is a viable endeavor to pursue.

Findings from the qualitative analysis of text-based survey responses show that students, in general, recounted positive experiences related to *Success* and *Caring* with their instructors. Mean student ratings from survey items related to *Success* and *Caring*, as well as *Overall Effectiveness of the Instructor*, also indicate generally positive feedback. This, however, is not consistent with the narrative from exisiting literature about the learning experience in large classes [13], [22], [40]. They are also not consistent with student performance in Mechanics courses, especially in Statics, where a large number of students are unsuccessful in the course on their first attempt [39]. These results may imply the following:

a) Students recognize and appreciate conscious effort on the part of Mechanics instructors to create an effective learning environment, despite the challenges posed by the large class setting. They do not seem to automatically associate the challenging nature of the courses and experiences of frustration and failure when going through course material with the effectiveness of the instructor.

   This implication, however, may be attributed to, and limited by, the nature of participation in SPOT surveys. Since students self-select into the process, there may be a tendency for skewed results from respondents who are on either end of two possible extremes: students who are very pleased with their instructor and are enthusiastic about heaping praise, or students who have been terribly upset and use the survey as an opportunity to express their frustrations. Also, since the surveys are conducted toward the end of the semester, these do not reflect the sentiments of students who have already dropped the course early in the semester.

b) While there are coded instances of strategies and behaviors exhibited by instructors that need improvement, these are not as prevalent across instructors. Statements that express a need for improvement or changes in the way learning is facilitated constitute approximately 25% of the total coded statements; the percentage is even lower for the *Caring* dimension (11%). These statements are also clustered around specific instructors. This tendency may also be related to the nature of participation in the survey as discussed in the section above.

c) The current survey instrument may not sufficiently capture the range of student experiences in the large class setting. This implication is being explored due to the discrepancy between the findings that emerged from this study and literature on the educational environment in large classes [13], [13], [14]. While it is encouraging to find that students share positive experiences in their responses, we may need to consider whether the survey items are giving them enough prompting and opportunity to describe their experiences completely.

These implications may be verified and validated by looking further in to the learning experience, such as conducting classroom observations and interviews with students. It may also be helpful examine the current SPOT survey instrument to see whether making changes to current language or adding specific survey items are options worth pursuing.

As mentioned in the *results* and *key findings* sections, the use of natural language processing techniques to qualitatively analyze text responses to survey items seems promising. Additional runs to train the classification algorithm using manually-coded data from the remaining semesters (Spring 2015, Fall 2015, and Spring 2016) may improve accuracy; provide more baseline keywords, sample statements for classifying responses; and prepare the algorithm for more specific processing (such as distinguishing whether a statement is a positive or needs improvement instance of a construct). Future work will include manually coding data from the remaining semesters, starting initially with the *Success* and *Caring* constructs. Once training runs have been conducted using coded data from the purposefully-selected sample (1,262 responses), a trial run analyzing the entire dataset of 3,917 responses for *Success* and *Caring* may be conducted and evaluated for accuracy. The same process may then be followed for the other components of the MUSIC Model (*eMpowerment, usefulness,* and *interest*). The iterative process of manual qualitative analysis and conducting trial and training runs on the algorithm will ensure that the highest possible accuracy will be achieved. It is imperative for the algorithm to be as accurate as possible in order to ensure the integrity of results generated by an automated process.

In general, qualitatively analyzing the answers to open-ended survey items allowed us to consolidate and generalize, to a reasonable extent, student perceptions of their learning experiences in the courses that were analyzed. We specifically were able to identify student preferences for what constitutes a positive learning experience, and how these preferences varied across respondents, instructors, and courses. Further work will ultimately lead to a structure of disseminating text-based survey responses concisely, and in a form that presents feedback constructively, using automated text analysis.

REFERENCES

[1] National Science Board, "Science and Engineering Indicators 2014," National Science Foundation, 2014.

[2] M. Parry, "'Supersizing' the College Classroom: How One Instructor Teaches 2,670 Students," *The Chronicle of Higher Education*, 29-Apr-2012.

[3] J. Cuseo, "The empirical case against large class size: adverse effects on the teaching, learning, and retention of first-year students," *J. Fac. Dev.*, vol. 21, no. 1, pp. 5–21, 2007.

[4] J. C. Chen, D. C. Whittinghill, and J. A. Kadlowec, "Classes That Click: Fast, Rich Feedback to Enhance Student Learning and Satisfaction," *J. Eng. Educ.*, vol. 99, no. 2, pp. 159–168, Apr. 2010.

[5] Gallogly College of Engineering, "Degree Requirement Checksheets." University of Oklahoma, 2015.

[6] "Engineering | Embry-Riddle Aeronautical University," 2016. [Online]. Available: https://erau.edu/degrees/engineering/. [Accessed: 22-Mar-2017].

[7] "Rose-Hulman - Top Ranked Engineering College," 2016. [Online]. Available: http://www.rose-hulman.edu/. [Accessed: 22-Mar-2017].

[8] "Smith College: Picker Engineering Program," 2016. [Online]. Available: https://www.smith.edu/engineering/. [Accessed: 22-Mar-2017].

[9] "Virginia Tech College of Engineering | College of Engineering," 2016. [Online]. Available: https://www.eng.vt.edu/. [Accessed: 22-Mar-2017].

[10] University of Michigan, "Michigan Engineering." 2015.

[11] K. L. Coburn and M. L. Treeger, *Letting go: A parents' guide to understanding the college years*, 3rd ed. New York, NY: HarperCollins Publishers, Inc., 1997.

[12] E. Seymour and N. M. Hewitt, *Talking about leaving: why undergraduates leave the sciences*. Westview Press, 1997.

[13] E. Carbone and J. Greenberg, "Teaching Large Classes: Unpacking the Problem and Responding Creatively," in *To Improve the Academy*, vol. 17, M. Kaplan, Ed. Stillwater, OK: New Forums Press and the Professional and Organizational Development Network in Higher Education, 1998, pp. 311–326.

[14] C. Mulryan-Kyne, "Teaching large classes at college and university level: challenges and opportunities," *Teach. High. Educ.*, vol. 15, no. 2, pp. 175–185, Apr. 2010.

[15] M. Soledad and J. Grohs, "Understanding Faculty Experiences in Teaching Large Classes: A Pilot Study on Faculty Perceptions of Teacher-Student Interaction in Foundational Engineering Courses," in *The 2nd Annual Teaching Large Classes Conference*, 2016.

[16] J. Turns, J. Yellin, Y.-M. Huang, and B. Sattler, "We All Take Learners Into Account In Our Teaching Decisions: Wait, Do We?," presented at the 2008 Annual Conference & Exposition, 2008, p. 13.1391.1-13.1391.19.

[17] R. B. Barr and J. Tagg, "From Teaching to Learning: A New Paradigm for Undergraduate Education," *Change*, vol. 27, no. 6, pp. 12–25, Nov. 1995.

[18] E. Blair and K. Valdez Noel, "Improving higher education practice through student evaluation systems: is the student voice being heard?," *Assess. Eval. High. Educ.*, vol. 39, no. 7, pp. 879–894, Nov. 2014.

[19] H. W. |Roche Marsh, "Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility," *Am. Psychol.*, vol. 52, no. 11, pp. 1187–97, 1997.

[20] N. P. Pitterson, S. Brown, K. A. Villanueva, and A. Sitomer, "Investigating current approaches to assessing teaching evaluation in engineering departments," in *Frontiers in Education Conference (FIE), 2016 IEEE*, 2016, pp. 1–7.

[21] N. Punyanunt-Carter and S. L. Carter, "Students' Gender Bias in Teaching Evaluations," *High. Learn. Res. Commun.*, vol. 5, no. 3, p. 28, Sep. 2015.

[22] J. Turns, M. Eliot, R. Neal, and A. Linse, "Investigating the Teaching Concerns of Engineering Educators," *J. Eng. Educ. Wash.*, vol. 96, no. 4, pp. 295–308, Oct. 2007.

[23] M. Soledad, J. R. Grohs, J. Williams, S. Culver, and J. Doggett, "Leveraging Institutional Data for Reflective Teaching in Large Classes," presented at the 9th Annual Conference on Higher Education Pedagogy, Blacksburg, VA, 2017.

[24] B. D. Jones, "Motivating Students to Engage in Learning: The MUSIC Model of Academic Motivation.," *Int. J. Teach. Learn. High. Educ.*, vol. 21, no. 2, pp. 272–285, 2009.

[25] B. D. Jones, *Motivating students by design*. publisher not identifed, 2015.

[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, vol. 26. Upper Saddle River, NJ: Prentice Hall, 2007.

[27] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959.

[28] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide Note Classification Using Natural Language Processing: A Content Analysis," *Biomed. Inform. Insights*, vol. 2010, no. 3, pp. 19–28, Aug. 2010.

[29] J. Savoy, "Text clustering: An application with the State of the Union addresses," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 8, pp. 1645–1654, 2015.

[30] M. Soledad, J. Grohs, J. Williams, J. Doggett, and S. Culver, "Student Perceptions of Learning Experiences in Large Mechanics Classes: An Analysis of Student Responses to Course Evaluation Surveys," in *Conference Proceedings of the 2017 American Society for Engineering Education Annual Conference*, Columbus, Ohio, Accepted.

[31] H. G. Cheng and M. R. Phillips, "Secondary analysis of existing data: opportunities and implementation," *Shanghai Arch. Psychiatry*, vol. 26, no. 6, p. 371, Dec. 2014.

[32] K. A. Neuendorf, *The Content Analysis Guidebook*. Sage Publications, 2002.

[33] H.-F. Hsieh and S. E. Shannon, "Three approaches to qualitative content analysis," *Qual. Health Res.*, vol. 15, no. 9, pp. 1277–1288, 2005.

[34] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.

[35] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[36] J. Saldana, *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2009.

[37] E. Gabrilovich and S. Markovitch, "Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge," in *AAAI*, 2006, vol. 6, pp. 1301–1306.

[38] G. W. Ellis, A. N. Rudnitsky, and G. E. Scordilis, "Finding Meaning in the Classroom: Learner-Centered Approaches that Engage Students in Engineering," *Int. J. Eng. Educ.*, vol. 21, no. 6, pp. 1148–1158, 2005.

[39] J. Grohs, M. Soledad, D. Knight, and S. Case, "Understanding the Effects of Transferring In Statics Credit on Performance in Future Mechanics Courses," 2016.

[40] E. Carbone, "Students Behaving Badly in Large Classes," *New Dir. Teach. Learn.*, vol. 1999, no. 77, pp. 35–43, 1999.