2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #32993

Using Natural Language Processing to Facilitate Student Feedback Analysis

Dr. Andrew Katz, Virginia Polytechnic Institute and State University

Andrew Katz is an assistant professor in the Department of Engineering Education at Virginia Tech. He leads the Improving Decisions in Engineering Education Agents and Systems (IDEEAS) Lab, a group that uses multi-modal data to characterize, understand, and improve decisions made throughout engineering education ecosystems.

Matthew Norris, Virginia Polytechnic Institute and State University

Matthew Norris is a PhD student and Graduate Research Assistant in the Department of Engineering Education at Virginia Tech.

Abdulrahman M. Alsharif, Virginia Polytechnic Institute and State University

Abdulrahman M. Alsharif is a Ph.D. student and a research assistant in the Engineering Education department at Virginia Tech. He has received the Saudi Arabia Ministry of Higher Education and Scientific Research scholarship to pursue his Bachelor's and Master's degrees in Industrial and Systems Engineering. His research interests are teaching and learning, policy and guidelines, and assessments. He hopes to work as a social scientist in engineering in higher education.

Dr. Michelle D. Klopfer, Virginia Polytechnic Institute and State University
Dr. David B. Knight, Virginia Polytechnic Institute and State University

David B. Knight is an Associate Professor in the Department of Engineering Education and Special Assistant to the Dean for Strategic Plan Implementation at Virginia Tech. He is also Director of Research of the Academy for Global Engineering at Virginia Tech and is affiliate faculty with the Higher Education Program. His research tends to be at the macro-scale, focused on a systems-level perspective of how engineering education can become more effective, efficient, and inclusive, tends to leverage large-scale institutional, state, or national data sets, and considers the intersection between policy and organizational contexts. He has B.S., M.S., and M.U.E.P. degrees from the University of Virginia and a Ph.D. in Higher Education from Pennsylvania State University.

Dr. Jacob R. Grohs, Virginia Polytechnic Institute and State University

Jacob Grohs is an Assistant Professor in Engineering Education at Virginia Tech with Affiliate Faculty status in Biomedical Engineering and Mechanics and the Learning Sciences and Technologies at Virginia Tech. He holds degrees in Engineering Mechanics (BS, MS) and in Educational Psychology (MAEd, PhD).

Using Natural Language Processing to Facilitate Student Feedback Analysis

Abstract

This research paper compares the results of a novel computer-assisted approach for analyzing a large volume of open-ended responses with those of a more traditional open coding approach. The work is motivated by the observation that in engineering education ecosystems, community members produce text through myriad activities both inside and outside of the classroom in teaching and research settings. In many of these cases, there is an abundance of text available to educators and researchers that *could* provide insight into various phenomena of interest within the system - student conceptual understanding, student experiences outside the classroom, how instructors can improve their teaching, or even shifts in collective conversations. Unfortunately, while these bodies of text have the potential to provide novel insights to educators and researchers, traditional analysis techniques do not scale well. For example, analyzing larger amounts of text can take one grader or researcher significantly more time than grading a small set of text responses. A larger body of text also creates more challenges for intrarater reliability. Likewise, expanding the size of the grading or research team can create interrater reliability challenges and the possibility of bias.

To address this opportunity, we have created a natural language processing system that augments human analysis so as to facilitate and enhance the work of one person (or team). Specifically, we take minimally pre-processed text, embed them using a pre-trained transformer (a specific kind of neural network architecture trained to encode inputs and decode outputs), and perform a sequence of dimension reduction techniques capped with a final clustering step. Such a system can help reduce the amount of time needed to analyze the text by effectively running a first pass on the text to group similar responses together. The human user can utilize these groupings to perform further analysis to fine tune and identify meanings in ways that only a human could. The system also can help improve consistency by analyzing across the entire collection of texts simultaneously and grouping similar items together. This is in contrast with a single person or a team that would have to work in series, analyzing responses sequentially and thereby creating the potential for inconsistencies across time.

In this paper we describe the system's architecture and data processing steps. We demonstrate the utility of this approach by applying the method on three questions from an end-of-semester feedback survey in a large, required introductory engineering course. The survey questions were part of a general feedback survey and asked students about their experiences in the transition to online learning subsequent to the SARS-CoV-2 outbreak..

Our results suggest that the pre-analysis text clustering can improve speed and accuracy of coding when compared with unassisted human coding—the system augments what we have traditionally done in coding, grading, or making sense of large quantities of textual data. As natural language processing techniques continue to develop, the engineering education research community should continue to explore potential applications to improve understanding and sensemaking from large volumes of underutilized text data from both within and outside of classroom settings.

1. Introduction

Although engineering may be a discipline that employs mathematical and scientific concepts to design solutions to problems, natural language – emergent, human language [1] – can be found throughout engineering education. Indeed, one might encounter significant challenges in trying to teach and research without natural languages. In practice, this means that there exist ample opportunities to learn about the professional formation of engineers by looking at language used throughout the engineering education ecosystem. However, difficulties in scaling traditional qualitative research methods to analyze large amounts of language present an obstacle. For example, it can take one person a significant amount of time to read thousands of responses to a question, and as they analyze a corpus issues of consistency start to emerge. One solution: divide the work amongst a team, but the issue of total number of human-hours of work needed for the analysis persists. Moreover, that strategy can raise another issue of consistency not only across time but also across individuals. Software tools like NVivo and ATLAS.ti can help to address some of these issues when it comes to work organization and management across team members, but there are still ample areas for improvement. We suggest that natural language processing (NLP) presents a promising alternative solution.

Natural language processing is a term to describe a range of techniques for analyzing natural human languages [2], [3]. Some of the original NLP techniques involved rule-based grammars and analysis on adjacency matrices [4]. In 2003, Blei et al. [5] introduced a probabilistic graphical modelling approach called Latent Dirichlet Allocation to analyze a corpus of text with thousands of documents in order to fit a distribution of words over topics and topics over documents. This was one popular example of a general approach to NLP called topic modeling. Other popular NLP tasks include translation, sentiment analysis, question answering, named entity recognition, and text classification [6].

In this paper, we are proposing a human-in-the-loop approach to assist thematic coding via a text classifier that clusters segments of text (e.g., responses to open-ended survey questions) together in order to help researchers' analyses of large ($N > 1{,}000$) text corpora in engineering education. The reader should note: while prototypical classification systems typically require labeled training examples to train a neural network under a supervised learning paradigm, we employed NLP developments in attention mechanisms [7] and transformer architectures [8] to circumvent the labor intensive step of labeling a training set for such a supervised approach. Instead, the system we discuss herein can achieve accurate and meaningful classifications without pre-made labels simply by leveraging pre-trained sentence transformers trained on a standard corpus of Wikipedia articles. This unsupervised approach to labeling the data that circumvents the extra hours training the systems is one reason that we believe the approach in this paper may be a useful contribution to the field of engineering education. The main focus of this paper is to describe our general approach and offer a comparison of results from the human-in-the-loop system's performance with the results of a more traditional qualitative coding approach.

2. Prior Work

The advent of neural networks and hardware to support their training has introduced a new era for NLP research and applications. For example, these machine learning architectures helped introduce ideas like word2vec [9] and global vectors (GloVe) [10]. Informed by the notion of a distributional representation and the linguistic slogan "you shall know a word by the company it keeps" [11], these systems – and many subsequent systems – seek to represent a word or sequence of words in a high-dimensional (e.g., 100 or 300 dimensions) vector space. A neural network can then use these vectors as inputs for training and testing the network or any other array of activities. From there, architectures like convolutional neural networks and recurrent neural networks were trained for various NLP tasks. Notably, each of these architectures had their own limitations. To address these issues, the idea of attention was introduced in which a neural network can differentially attend to certain pieces of input [7]. When combined with an autoencoder, this eventually culminated in transformer architectures. These transformer-based models are the backbone of the system we have used for this research. In addition to being state-of-the-art in many NLP systems, transformers can allow researchers to develop a system that does not require thousands of additional labeled training examples. Instead, they are trained on billions of tokens of text in a self-supervised manner to generate a probabilistic language model that can then be used for downstream tasks such as ours. This possibility makes transformers an advantageous element for a system such as ours because they meet the design requirement that the system should need minimal additional training. Satisfying the "minimal additional training" requirement means that the engineering education teaching and research community could apply the this kind of system off the shelf in their own work to identify important trends and answer relevant questions in their own contexts.

In educational data, NLP techniques have been used to study a variety of topics. Crossley et al., [12], [13] used a series of rule-based approaches to study students' sentiments and their math identities in an intelligent tutoring system. Crossley et al [14] also used an NLP approach to study differences in students writing styles as a function of their disciplines. A third example involves classifying the quality of questions that students generated when using an English writing intelligent tutoring system, once again using a rule-based system [15]. In the area of analyzing feedback surveys, Dhanalakshmi et al. [16] used a supervising learning approach to predict the polarity of student responses (a common framing of a sentiment analysis task). Of course, these models also have several potential limitations such as inadvertently introducing bias and reflecting unintentional differences across groups [17], [18].

In engineering education, there have been limited applications of NLP on either the research or teaching side. The more modern applications have applied standard statistical and machine learning techniques such as rule-based classifiers for assessing student responses [19]–[21], college mission statements [22], writing exercises [23], and emotions in student stories of their transitions to university [24]. Unfortunately, these kinds of rule-based systems tend to be brittle and poorly handle variations in language to express the same concept. For example, those systems may require common word usage across responses in order to recognize that they are expressing the same idea. As such, it would fail to identify the similarity between student A responding "I didn't have to get dressed up for class" and student B responding "We could wear pajamas from the comfort of our homes" in response to a question like "What did you like about the remote lab setup?" (an actual example from [25]). Even though those responses share no words in common, they express the same idea. A more flexible approach such as that presented

by neural networks can help address this problem. The underutilization of neural network-based NLP systems highlights how much room there could be for a system like the one presented here to transform how teachers and researchers utilize texts generated during the formation of engineers. To that end, we sought to answer the following research question in this work: *how can a computer-assisted coding approach using modern NLP techniques perform when compared with a more traditional coding approach for analyzing student survey responses?*

3. Methods

In this section we describe both the data collection and data analysis methods used to address this research question.

3.1 Data Collection

We collected data through a series of questions included in a course feedback survey. The course was a first-year engineering course in Spring 2020 with more than 25 sections. Near the middle of the semester, the university transitioned all classes to an online modality. As a result of the shift, the survey contained a series of closed and open-ended questions beyond the typical questions on the survey in order to glean information about student experiences with the transition. Student responses to individual survey items were optional. For this study, we used three of the open-ended questions. Table 1 lists the questions and the number of responses since each question was optional.

Table 1. Questions from end-of-semester survey used in analysis

Question	No. of Responses
What did your [course] instructor do during the transition to online learning that helped you to stay engaged in the course?	1,197
What could your [course] instructor have done differently in the online transition to help you learn?	1,066
If we enter this situation again, what would you recommend to a peer to be successful?	1,212

3.2 Data Analysis

For this study, we analyzed the three questions with two approaches. The first was a more traditional qualitative coding approach. The second was a computer-assisted approach. This second approach is the novel contribution at the center of this particular paper.

3.2.1 Traditional approach

To analyze responses to survey questions, we used an open coding approach to identify emergent themes from available responses similar to thematic analysis [30], [31]. The process starts by manually looking across the first 100 responses and identifying common themes. Coders then

began coding survey responses, adding codes any time responses were identified to fall in categories not covered by preexisting codes. Individual responses with more than one idea represented in the text were assigned multiple codes. Codes were collapsed into higher level concepts when categories were similar and had few responses. Distinct themes were retained even in the case of only having a small number of cases.

3.2.2 Computer-assisted approach

To analyze these survey questions, we used a novel combination of modern natural language processing techniques. The process starts with the raw text from student responses. We then embed the sentences in a high dimensional vector space with sentence transformers based on the BERT architecture [26]. These embeddings then undergo a combination of linear and nonlinear dimensionality reduction steps using principal component analysis (PCA) and uniform manifold approximation and project (UMAP) [27], respectively. We used PCA to reduce from the original embedding space into an intermediate embedding space since the technique is efficient at maintaining variance in the original embeddings without losing too much information. An example of the reduction in variance is shown in figure 1. From this original embedding space, we then reduced to a lower dimensional ($p < 5$) space using UMAP. In the lower dimensional space, clustering becomes more feasible. Clustering in the intermediate space historically suffers from a curse of dimensionality wherein every point (i.e., text embedding) is far from every other point [28].

Finally, to cluster the data, we tried three separate clustering algorithms - k-means, hierarchical density based spatial clustering for applications with noise (HDBSCAN) [29], and agglomerative clustering. We determined that agglomerative clustering produced the most internally homogeneous text groupings (meaning that texts clustered together most frequently discussed the same topic with the agglomerative clustering method in contrast with the other two). Moreover, with agglomerative clustering, we used ward linkage with a Euclidean distance metric because this produced the most consistent groupings when compared with other linkage options and metrics. With these cluster labels, a member of the research team then coded each cluster by reading the responses. This proceeded significantly quicker – on the order of a five-fold reduction in time – than the baseline coding since each cluster could contain in the range of 20-60 responses, and when the coder identified the theme in that entire group then they could assign one label to each of these responses and move on to the next cluster. We estimate that this approach leads to a significant reduction in time coding the data while enabling analysis of large volumes of data that were previously unwieldy to handle.

For comparison, two separate members of the research team also coded the original responses for each of the three questions using thematic analysis [30], [31]. This helped to serve as a baseline to determine areas for improvement with the NLP approach. The process was compared against this baseline in two dimensions: time to complete coding and accuracy. The three team members involved in this portion of the work did not confer during the coding, but they did meet to discuss their results after coding was completed.

4. Results

In this section we will first present the results for the NLP embedding and reduction steps in order to illustrate the tradeoff between enabling the clustering algorithm to identify meaningful clusters while not reducing the dimensions of the original embedding to the point where valuable information is lost. Next, we will compare the results of the traditional qualitative coding in contrast with the computer-assisted approach. The reader should note that one significant difference in the two approaches was the time required for each coding approach. The traditional qualitative coding approach took four to five hours per question whereas the computer-assisted approach took one hour to complete. We believe this might be one of the more useful results of this study since it suggests an ability to scale analysis of certain kinds of qualitative data.

4.1 Dimension Reduction

For the dimension reduction step, we first used PCA to reduce from the original raw text embedding space (e.g., 1,024 dimensional space) to an intermediate embedding space (in the range of 50-100 dimensions). The goal here was to reduce to a space where a nonlinear technique can reasonably function (e.g., less than 100 dimensions). Unfortunately, PCA can result in losing valuable information in the process of dimension reduction if there is too sharp of a reduction. This requires balancing the tradeoff between maintaining as much information in the original data as possible while still enabling the UMAP step to work well. Figure 1 illustrates this tradeoff, showing how the preserved variation in the original data decreases as the PCA projection dimension decreases from 80 dimensions to 0 dimensions. Experimentally, we found that an intermediate embedding in the range of 65-80 dimensions appears to balance these two considerations best - maintaining in the range of 75-90% of the original variance in the data while still enabling UMAP to function properly.

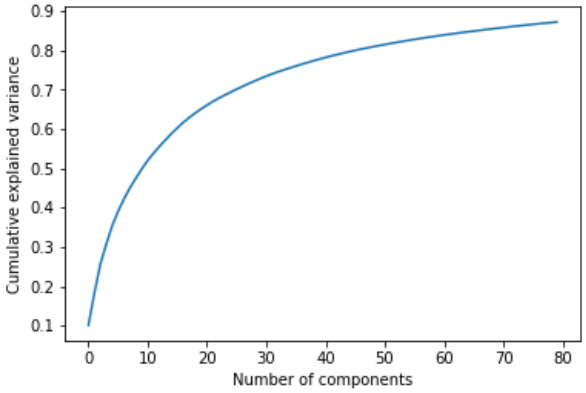

Figure 1. Tradeoff associated with PCA to balance dimensions of projection (number of components) and amount of variance explained by dimensions (cumulative variance explained).

4.2 Item 1: What instructors did to maintain student engagement.

The first question was about what the course instructor did that the students said helped keep them engaged. The traditional coding approach identified 26 different topics, as shown in Figure 2. In comparison, the computer-assisted approach identified 25 different topics. In each approach, material posted by instructors in the form of lecture videos, weekly update videos, clear descriptions of assignments, and office hours were some of the most frequently cited things that helped keep students engaged. There were also several types of communication either from

check-ins, more office hours, online discussions, and generally good communication that were identified. Each approach also identified a group of students saying that nothing kept them engaged. The computer-assisted approach also identified a group of students who specified the instructor's humor as keeping them engaged, which the traditional approach did not identify. When the analysts from the two approaches compared their results, they believe this was partially due to some of the humor being subsumed into the "enjoyable material" category for the traditional approach. The overlap and differences of each approach are shown in Figure 2. A common theme in each of these figures for each question was the relative scale of the counts. The computer-assisted approach (in its current implementation, as highlighted in the limitations section) was designed for assigning one label per student response. In contrast, the traditional, thematic coding had no such restriction. This means that the total number of codes in the computer-assisted approach had a cap of 1,197 labels since there were 1,197 responses to this question. In contrast, the traditional coding approach was theoretically unlimited in the number of labels appended to responses since those responses could each have multiple codes applied.

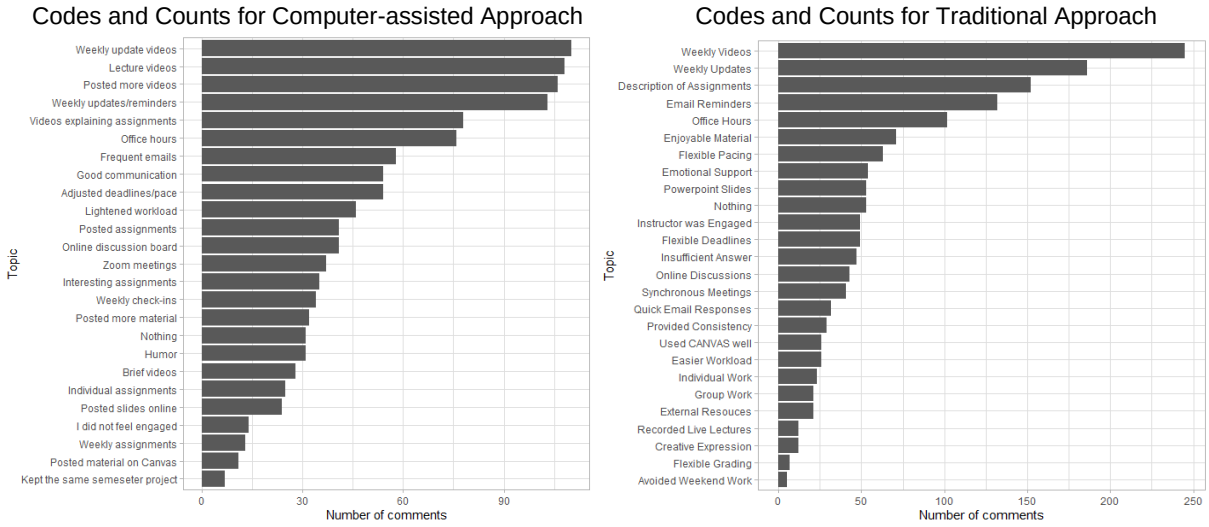

Figure 2. Distribution of topics and their respective counts from coding using (a) computer-assisted approach or (b) traditional qualitative coding approach for question 1, What did your [course] instructor do during the transition to online learning that helped you to stay engaged in the course?

4.3 Item 2: Where instructors could have improved.

The second question asked about what the instructor could have done to help students learn. The traditional coding approach identified 19 topics. The computer-assisted approach identified 13 topics. In each approach, the vast majority of respondents either said the instructor could not have done anything differently or went one step further and said the instructor did a great job. Unlike with other items, the counts for these two were roughly similar. This was most likely because students' responses expressing these two topics were relatively short and only expressed this one idea; hence, the design limitation of computer-assisted approach only being able to append one label to a response was not a hindrance. Of the more constructive responses, each approach suggested clearer communication, more communication, more live classes, and less

teamwork. One notable discrepancy between the two approaches was the identification of how instructors were using Canvas. This did not show up in the computer-assisted approach. We believe this is because comments about Canvas were clustered with a larger group of comments about either communication (as in more communication through Canvas) or clearer instructions (as in clearer instructions provided on Canvas). Class participation also did not show up in the computer-assisted approach. We believe some of these comments were included in the live classes and more online meetings topics, although that does not account for the entire discrepancy between the two approaches. An alternative explanation for this is that comments may have mentioned more class participation along with another suggestion and that suggestion was the one driving the labeling with the computer-assisted approach. This underscores a limitation to the current implementation that we expound upon in the discussion. An important takeaway is the value of considering responses at the sentence level rather than taking entire responses and clustering those in their entirety—our future work will investigate that level of analysis.

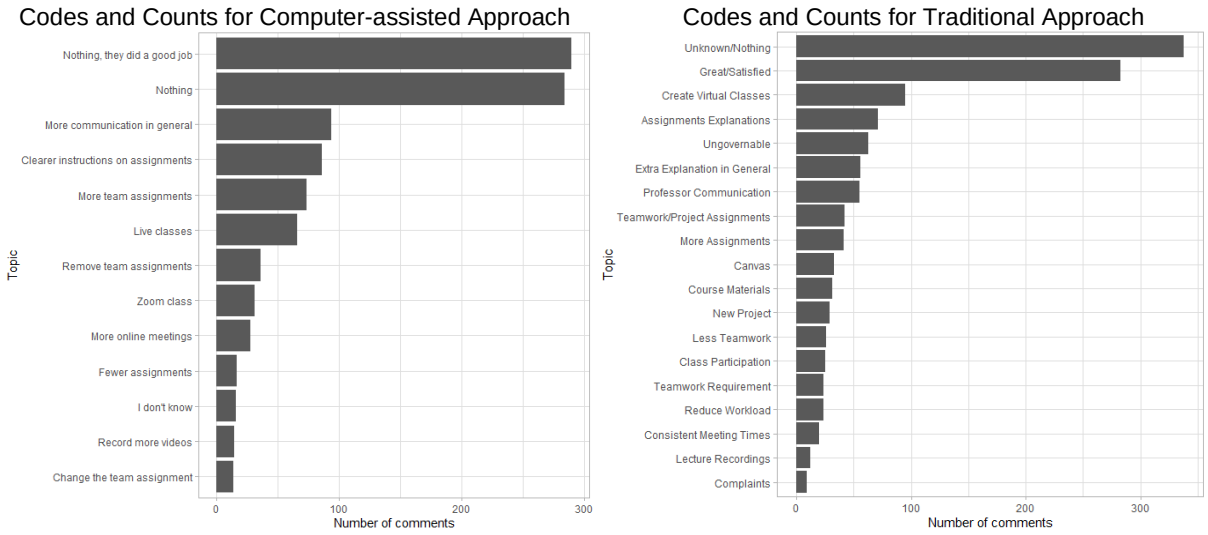

Figure 3. Distribution of topics and their respective counts from coding using (a) computer-assisted approach or (b) traditional qualitative coding approach for question 2, What could the instructor have done to help you the instructor do to help you as a learner?

4.4 Item 3: Students' suggestions to future students

Whereas the first two questions focused on the instructor's actions, the third question asked students what they would suggest to someone in their own role, as a student, to be successful if this scenario were to happen again. The traditional qualitative coding approach identified 19 different topics. The computer-assisted approach identified 21 different topics. One of those topics was labeled as "noisy" because there were too many different concepts expressed in each of those responses and assigning a single label was unrealistic. A second cluster was labeled as "noise" because this set of responses was not meaningful (i.e., responses that said "yes" to this question, even though it was not a yes/no question). There was also a large discrepancy in the number of labels between the two approaches because the traditional approach permitted multiple labels assigned to a single response while the computer-assisted approach (in its current

implementation) only permitted one label per response. Future versions will allow more flexibility in order to match the traditional approach and permit multiple labels (i.e., codes) to be applied to a participant's response.

The substance of the labels was a little more divergent for this question compared to the previous two questions. For an example of overlap in each approach students suggested some combination of work-related ideas like not falling behind, not procrastinating, working ahead, and sticking to a schedule. There was also a collection of comments about communicating with teammates (the course typically involves a significant amount of teamwork). There were also several health (physical and mental) related suggestions like keeping a regular schedule and getting enough sleep. Each approach also identified a non-trivial number of suggestions that students drop out for the semester. On the other hand, the computer-assisted approach identified a large number of comments suggesting that students check Canvas regularly. Some of those comments were tagged with the traditional approach as being about requirement awareness or workload awareness (as in, check Canvas to be aware of assignments). Another larger discrepancy was in how many comments were tagged as health awareness. After the two teams compared their responses, some of this seemed to be from the computer-assisted approach separating out different aspects of health awareness (e.g., maintaining regular sleep) from the general umbrella code of health awareness.

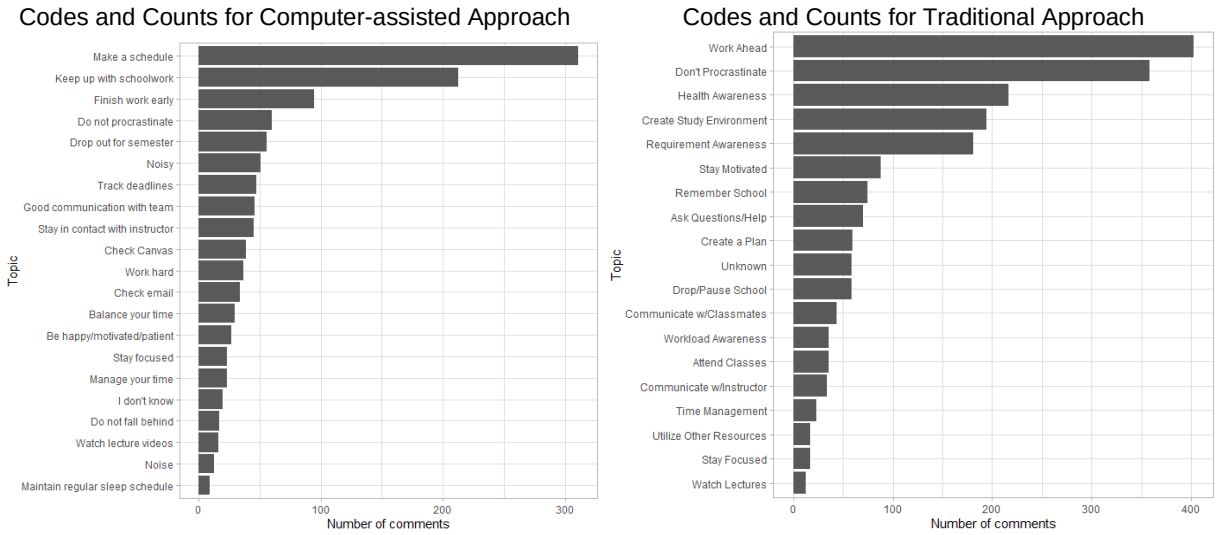

Figure 4. Distribution of topics and their respective counts from coding using (a) computer-assisted approach or (b) traditional qualitative coding approach for question 3, If we enter this situation again, what would you recommend to a peer to be successful?

5. Discussion

This study highlighted several important contributions and drawbacks from incorporating NLP to analyze responses such as these. As mentioned above, this kind of technique is superfluous when working with small numbers of observations. The purpose is not to replace the human qualitative researcher, but rather to enable them to work with larger volumes of data that previously presented challenges of time and coordination across a team of researchers. For example, in this work, it reduced the amount of time and number of people needed to analyze > 3,000 survey

responses from students. In the traditional approach, this coding took ~12 hours to code while it took ~ 3 hours with the computer-assisted approach. As the number of responses increases, this advantage becomes even more pronounced. In other work under preparation, for example, it has taken ~ 6 hours to code > 200,000 responses to surveys collected over a period of 10 years with the computer-assisted approach. Working on that scale of data can more readily translate into identifying systematic patterns in the data that could combine some of the features of qualitative research and quantitative research together in order to answer previously difficult-to-answer research questions.

In another point of comparison, these two approaches both identified similar themes in each of the three questions. To this end, even if there were discrepancies between the overall counts of themes between the approaches, this computer-assisted approach could help with either building a codebook or as a check of an existing codebook. When the three members of the team involved in the analysis (two from the traditional approach and one from the computer-assisted approach) compared their codes, they noticed multiple similarities in the kinds of codes they identified. In some instances, different words were used to signify the same underlying idea. For example, "finish work early" and "work ahead" or "communicate w/ instructor" and "stay in contact w/ instructor" were coded with different language in the two approaches even though they ostensibly mean the same thing. This helps explain some of the differences in the results shown above insofar as they look different because the words for each code do not match one-to-one. If the same person were labeling the traditional approach and the computer-assisted approach, we anticipate these differences would diminish considerably.

At this point, there are several notable limitations to emphasize. First, the computer-assisted approach stumbles when there are several topics discussed within one response. For example, if a student says, "I really enjoyed the short weekly update videos" then that presents no issues; however, if a student says, "I really enjoyed the short weekly update videos. The team assignments also really helped." then that presented a greater challenge. In future work we plan to split responses by sentence since the vast majority of sentences only expressed one topic. This would help mitigate the issue since the challenging responses were those that contained multiple sentences disconnected from each other. Analysis at the per-sentence level with the computer-assisted approach would also enable more comparisons between the two approaches since they would have a closer number of codes attached to the responses.

A second limitation is that implementing the system required a brief search over the parameter space to fine-tune the hyperparameters in the NLP model. At this early stage in the system's development, this required manual adjustment in the codebase rather than running through a convenient application programming interface. Future developments will include building an API so that the broader research community could use this approach without a familiarity with coding. This is part of our ultimate goal: provide the research community with accessible tools to assist their sensemaking of large-scale qualitative data. With that said, it is important to note that we view this approach in its current form as a way to facilitate work with qualitative data and as a way to potentially bridge between qualitative and quantitative analysis rather than favoring one approach over the other. We are taking a pragmatist perspective to finding novel ways to work with qualitative data at scale that enable the engineering education community to answer important questions (as defined by myriad communities).

A third limitation is the potential for unintended bias. In future work, we plan to assess this approach for potential unintended biases that minimize or completely elide comments from smaller groups of students. While the approach seems to handle grammatical issues and atypical sentence structures well, we plan to more systematically assess how it handles students for whom English is a second language. There were minimal examples of this in the current dataset, so we were not able to assess this aspect for the present study.

6. Conclusion

This approach is surely not necessary for small collections of qualitative data; rather, at this point, it has the most to offer in scenarios where faculty, administrators, and researchers have thousands of responses to targeted questions from which they want to glean information. While coding that quantity of data might take someone tens of hours to analyze, the approach discussed here can reduce that time to only a couple of hours. Moreover, while that volume of data could be split among a team of coders, this approach can be managed by a single coder.

Scaling qualitative analysis to that volume of data can facilitate identifying systematic variations in things like students' experiences during a pandemic when paired with other quantitative data such as responses to close-ended survey items. For example, researchers could approach sample sizes large enough to facilitate more traditional statistical modeling. Although that kind of analysis is beyond the scope of this paper, future work will pursue this line of questioning using student feedback surveys in order to identify potential systematic biases associated with characteristics of faculty members, students, and courses. This approach could also be helpful when integrated into some assessment scenarios where students provide open-ended responses and grading proceeds through a standard rubric. Identifying similar responses in those settings could potentially address inconsistencies in settings where grading was handled by a team and may have inadvertently introduced inconsistencies in the process. We also plan to use this approach with administrative data collected by the College of Engineering in order to illustrate how this could be utilized at an even larger scale to glean information from end-of-semester student feedback surveys. While these surveys have documented biases associated with them, they can also provide information about other aspects of students' experiences and do therefore can offer insights to departmental and college administrators when cabined appropriately.

References

[1] T. Winograd, "Understanding natural language," *Cognitive Psychology*, vol. 3, no. 1, pp. 1–191, Jan. 1972, doi: 10.1016/0010-0285(72)90002-3.

[2] K. R. Chowdhary, "Natural Language Processing," in *Fundamentals of Artificial Intelligence*, K. R. Chowdhary, Ed. New Delhi: Springer India, 2020, pp. 603–649.

[3] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[4] E. Brill and R. J. Mooney, "An Overview of Empirical Natural Language Processing," *AI Magazine*, vol. 18, no. 4, Art. no. 4, Dec. 1997, doi: 10.1609/aimag.v18i4.1318.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[6] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, May 2014, doi: 10.1109/MCI.2014.2307227.

[7] A. Vaswani *et al.*, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.
[8] C. Wang, M. Li, and A. J. Smola, "Language Models with Transformers," *arXiv:1904.09408 [cs]*, Oct. 2019, Accessed: Nov. 08, 2020. [Online]. Available: http://arxiv.org/abs/1904.09408.
[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv:1301.3781 [cs]*, Sep. 2013, Accessed: Nov. 06, 2020. [Online]. Available: http://arxiv.org/abs/1301.3781.
[10] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
[11] J. Firth, *A synopsis of linguistic analysis*. Oxford, UK: Blackwell, 1957.
[12] S. Crossley, J. Ocumpaugh, M. Labrum, F. Bradfield, M. Dascalu, and R. S. Baker, "Modeling math identity and math success through sentiment analysis and linguistic features," Jul. 2018, Accessed: Mar. 07, 2021. [Online]. Available: https://eric.ed.gov/?id=ED593117.
[13] S. A. Crossley, S. Karumbaiah, J. Ocumpaugh, M. J. Labrum, and R. S. Baker, "Predicting math identity through language and click-stream patterns in a blended learning mathematics program for elementary students," *Learning Analytics*, vol. 7, no. 1, Mar. 2020, doi: 10.18608/jla.2020.71.3.
[14] S. Crossley, D. Russell, K. Kyle, and U. Romer, "Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields?," *Journal of Writing Analytics*, vol. 1, pp. 48–81, Jan. 2017.
[15] K. J. Kopp, A. M. Johnson, S. A. Crossley, and D. S. McNamara, "Assessing question quality using NLP," in *Artificial Intelligence in Education*, Cham, 2017, pp. 523–527, doi: 10.1007/978-3-319-61425-0_55.
[16] V. Dhanalakshmi, D. Bino, and A. M. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms," in *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, Mar. 2016, pp. 1–5, doi: 10.1109/ICBDSC.2016.7460390.
[17] J. Vig *et al.*, "Investigating gender bias in language models using causal mediation analysis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12388–12401, 2020.
[18] N. Arthurs and A. J. Alvero, "Whose truth is the 'ground truth'? College admissions essays and bias in word vector evaluation methods," Jul. 2020, Accessed: Mar. 07, 2021. [Online]. Available: https://eric.ed.gov/?id=ED608067.
[19] C. A. Arbogast and D. Montfort, "Applying Natural Language Processing Techniques to an Assessment of Student Conceptual Understanding," presented at the 2016 ASEE Annual Conference & Exposition, Jun. 2016, Accessed: Nov. 08, 2020. [Online]. Available: https://peer.asee.org/applying-natural-language-processing-techniques-to-an-assessment-of-student-conceptual-understanding.
[20] S. S. Haris and N. Omar, "A rule-based approach in Bloom's Taxonomy question classification through natural language processing," in *2012 7th International Conference on Computing and Convergence Technology (ICCCT)*, Dec. 2012, pp. 410–414.
[21] M. A. Verleger, "Using Natural Language Processing Tools to Classify Student Responses to Open-Ended Engineering Problems in Large Classes," Jun. 2014, p. 24.1338.1-24.1338.15, Accessed: Nov. 08, 2020. [Online]. Available: https://peer.asee.org/using-natural-language-processing-tools-to-classify-student-responses-to-open-ended-engineering-problems-in-large-classes.
[22] S. Bhaduri and T. Roy, "Demonstrating use of natural language processing to compare college of engineering mission statements," presented at the 2017 ASEE Annual Conference & Exposition, Jun. 2017, Accessed: Nov. 08, 2020. [Online]. Available: https://peer.asee.org/demonstrating-use-of-natural-language-processing-to-compare-college-of-engineering-mission-statements.
[23] C. G. P. Berdanier, E. Baker, W. Wang, and C. McComb, "Opportunities for natural Llnguage processing in qualitative engineering education research: Two examples," in *2018 IEEE Frontiers in Education Conference (FIE)*, Oct. 2018, pp. 1–6, doi: 10.1109/FIE.2018.8658747.
[24] A. Satyanarayana, K. Goodlad, J. Sears, P. Kreniske, M. F. Diaz, and S. Cheng, "Using natural language processing tools on individual stories from First-year students to summarize emotions, sentiments, and concerns of transition from high school to college," presented at the 2019 ASEE Annual Conference & Exposition, Jun. 2019, Accessed: Mar. 07, 2021. [Online]. Available: https://peer.asee.org/using-natural-language-processing-tools-on-individual-stories-from-first-year-students-to-summarize-emotions-sentiments-and-concerns-of-transition-from-high-school-to-college.

[25] M. Szoke, A. Borgoltz, M. S. Kuester, N. Intaratep, W. J. Devenport, and A. Katz, “The Development of Remote Laboratory Sessions at the Stability Wind Tunnel of Virginia Tech During the Coronavirus Pandemic,” in *AIAA Scitech 2021 Forum*, American Institute of Aeronautics and Astronautics.
[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018, Accessed: Nov. 05, 2020. [Online]. Available: https://arxiv.org/abs/1810.04805v2.
[27] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv:1802.03426 [cs, stat]*, Sep. 2020, Accessed: Dec. 01, 2020. [Online]. Available: http://arxiv.org/abs/1802.03426.
[28] I. Assent, “Clustering high dimensional data,” *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 340–350, 2012, doi: 10.1002/widm.1062.
[29] L. McInnes, J. Healy, and S. Astels, “HDBSCAN: Hierarchical density based clustering,” *Journal of Open Source Software*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.
[30] V. Clarke and V. Braun, “Thematic analysis,” *The Journal of Positive Psychology*, vol. 12, no. 3, pp. 297–298, May 2017, doi: 10.1080/17439760.2016.1262613.
[31] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.
[32] M. D. Dixson, “Creating effective student engagement in online courses: What do students find engaging?,” *JoSoTL*, pp. 1–13, 2010.
[33] H. Peters *et al.*, “Twelve tips for enhancing student engagement,” *Medical Teacher*, vol. 41, no. 6, pp. 632–637, Jun. 2019, doi: 10.1080/0142159X.2018.1459530.
[34] J. Roksa, T. L. Trolian, C. Blaich, and K. Wise, “Facilitating academic performance in college: Understanding the role of clear and organized instruction,” *High Educ*, vol. 74, no. 2, pp. 283–300, Aug. 2017, doi: 10.1007/s10734-016-0048-2.
[35] X. Wang, S. Hegde, C. Son, B. Keller, A. Smith, and F. Sasangohar, “Investigating mental health of US college students during the COVID-19 pandemic: Cross-sectional survey study,” *Journal of Medical Internet Research*, vol. 22, no. 9, p. e22817, 2020, doi: 10.2196/22817.
[36] A. Kecojevic, C. H. Basch, M. Sullivan, and N. K. Davi, “The impact of the COVID-19 epidemic on mental health of undergraduate students in New Jersey, cross-sectional study,” *PLOS ONE*, vol. 15, no. 9, p. e0239696, Sep. 2020, doi: 10.1371/journal.pone.0239696.
[37] W. E. Copeland *et al.*, “Impact of COVID-19 pandemic on college student mental health and wellness,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 60, no. 1, pp. 134-141.e2, Jan. 2021, doi: 10.1016/j.jaac.2020.08.466.