2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering Education for 130 Years
ASEE
Paper ID #37764

Is Natural Language Processing Effective in Education Research? A case study in student perceptions of TA support

Neha Kardam, University of Washington

Neha Kardam is a third-year Ph.D. student in Electrical and Computer Engineering at the University of Washington, Seattle.

Ms. Shruti Misra, University of Washington

I am a graduate student in Electrical and Computer Engineering at the University of Washington, Seattle. My research interest is broadly focused on studying innovation in university-industry partnerships. I am interesting in various ways that universities

Dr. Denise Wilson, University of Washington

Denise Wilson is a professor of electrical engineering at the University of Washington, Seattle. Her research interests in engineering education focus on the role of self-efficacy, belonging, and other non-cognitive aspects of the student experience on e

Is Natural Language Processing Effective in Education Research?
A case study in student perceptions of TA support

Abstract

Natural language processing (NLP) techniques are widely used in linguistic analysis and have shown promising results in areas such as text summarization, text classification, autocorrection, chatbot conversation management, and many other applications. In education, NLP has primarily been applied to automated essay or open-ended question grading, semantic evaluation of student work, or the generation of feedback for intelligent tutoring-based student interaction. However, what is notably missing from NLP work to date is a robust automated framework for accurately analyzing text-based educational survey data. To address this gap, this case study uses NLP models to generate codes for thematic analysis of student needs for teaching assistant (TA) support and then compares code assignments for NLP vs. those assigned by an expert researcher.

Student responses to short answer questions regarding preferences for TA support were collected from an instructional support survey conducted in a broad range of electrical, computer, and mechanical engineering courses between 2016-2021 in engineering (N>1400) at a large public research institution. The resulting dataset was randomly split into training (60%), validation (20%), and test set (20%). A popular NLP topic modeling approach (Latent Dirichlet Allocation—LDA) was applied to the training dataset, which determined the optimal number of topics of code represented in the dataset to be four. These four topics were labeled as: (1) examples, where students expressed a need for TAs to illustrate additional problem-solving and applied content in engineering courses; (2) questions and answers, where students desired more opportunities to pose questions to TAs and obtain timely answers to those questions; (3) office hours, encompassing additional availability outside of formally scheduled class times; and (4) lab support. For the testing and validation datasets, an experienced researcher then used these four labels as codes to identify the ground truth for each student's response. Ground truth was then compared to NLP model predictions to gauge the accuracy of the model. For the validation dataset, the accuracy with which NLP identified each response as containing or not containing each code ranged from 79.4% to 91.1%, while for the testing dataset, such accuracies ranged from 81.1 to 92.2%. The codes identified by NLP were then combined into themes by a human researcher, resulting in three themes (problem-solving, interactions, and active/experiential learning). Conclusions reached regarding the three themes were identical whether the NLP codes or (human) researcher codes were used for data interpretation.

Short-answer questions, despite their value in providing deeper insight into the student experience, are infrequently used in educational research because the resulting data often requires prohibitive human resources to analyze. This study has demonstrated, in a case study of student preferences for TA support, the value of NLP in understanding large numbers of textual, short-answer responses from students. The fact that NLP models can deliver the same bottom line in minutes compared to the hours that traditional thematic analysis methods consume is promising for expanding the use of more nuanced, richer text-based data in survey-based education research.

Introduction

Natural language processing (NLP) is an interdisciplinary field that bridges concepts in linguistics, computer science, and artificial intelligence. NLP uses computers to preprocess, analyze, and interpret large amounts of natural language data (whether spoken or written) and is a growing field of study that aims to achieve human-like language processing for a wide range of tasks in an equally broad range of disciplines [1].

Among the disciplines that have benefitted from advances in NLP is education. NLP has been applied broadly in education spanning from education research to classroom teaching. NLP has been used to *assess* and *classify* student learning, to develop tools to *assist* in student learning, to analyze feedback *from students* and *about students*, and to understand feedback *from teachers*. Multiple approaches have been applied in these areas of education, some of which require the intervention of a human expert (human-in-the-loop, NLP-in-the-loop, human-as-expert) and some of which operate largely independent of a human expert (NLP-as-expert).

Regardless of how NLP has been integrated into teaching and research in education, it has almost universally been applied to substantial amounts of text-based data that would traditionally be analyzed using qualitative rather than quantitative techniques. Among those techniques, thematic analysis remains one of the most common techniques to understand and interpret such data and involves identifying and interpreting patterns of meaning within qualitative data that typically answer “how” or “why” questions regarding the data. In education, thematic analysis produces themes or codes that allow a deeper and richer understanding of teaching, student learning, or student experience. Thematic analysis provides insight into patterns and processes of behavior that are difficult if not impossible to gain using traditional quantitative methods of research [2].

Consistent with a great deal of qualitative research in education, this study uses *thematic analysis* to understand feedback *from students* regarding instructional (TA) support. It does so using NLP in conjunction with the assistance of a human expert (i.e., *human-in-the-loop*) to understand themes regarding what students want from TAs in engineering education settings.

Prior Use of NLP in Education

The use of NLP in education has been significant, particularly in the assessment and classification of student learning. Assessment involves determining the quality and level of student learning, while classification aims to comprehend student learning without evaluating it. Automated assessment is an attractive solution for large student populations, and one of the most common applications of NLP in education is the assessment of student writing in the Test of English as a Foreign Language (TOEFL) [3]. NLP is used to evaluate grammar, mechanics, word usage, complexity, style, and organization of student essays. NLP-based assessments have demonstrated remarkable agreement with teacher grades, ranging from 70% to 90%, when combined with neural networks [4]. In terms of vocabulary, NLP assessment tools have accounted for 44% of the variation in vocabulary knowledge among college students [5]. This success has led to the development of Automated Essay Grading (AEG) or scoring (AES) systems, some of which have been tailored to engineering education and are especially helpful for (English as a second language students) ESL students [6], [7].

While using NLP in assessment has valuable implications to substantially reducing grading overhead for teachers, it can also support teachers by breaking down or classifying what students are learning. For instance, NLP has been used to categorize student responses to a physics measurement problem by assigning student responses into one of three conceptual categories, with the same level of agreement as a human coder [8]. However, NLP was not able to perform finer classifications using subcodes employed by human graders [8]. In engineering, NLP has been applied to classify documents produced by student design teams, to predict the success of these teams, and to analyze student writing across various disciplines [9], [10], [11]. These efforts reinforce the notion that NLP is not only useful in producing a numerical or quantitative judgement of student work (i.e., a grade) but can also assist teachers in understanding differences among students and their learning.

Differences in students, however, are not limited to their written work nor is the relevance of NLP to education limited to the analysis of student work. NLP has also been successfully applied to the analysis of textual data provided by students (e.g., responses to open-ended questions on course surveys) to provide rich insight into differences in motivation [12], [13], identity [14], goals [15], and persistence [16]. In addition, NLP has been applied to sentiment analysis in education, especially to understand student emotions regarding their educational experiences. NLP approaches to sentiment analysis have achieved high accuracy levels, ranging from 75% to 99% when compared to traditional human approaches [17],[18]. More recently, sentiment analysis has been applied to a finer grained analysis of sentiment detecting tones of joy fear, sadness, anger, analytic, confident, and tentative in student generated stories of their lived experiences [19]. Student sentiment in feedback regarding their educational experiences is important, but it alone does not provide sufficient information to act on that feedback. To facilitate improvements in the educational experience, more information is needed, and NLP may play a more nuanced role in automating or augmenting such analysis by identifying codes or themes in thematic analysis of textual data [20], [21].

Approaches to the Use of NLP in Education Research

The application of NLP to education can be broken down by its function broadly into four main categories: (a) *human-as-expert* approaches which compare the results of NLP-based machine learning to a human researcher, presumed to be an expert; (b) *NLP-as-expert* studies which assume that results of the automated classification of text-based data are of value in and of themselves without the intervention or approval of a human expert; (c) *NLP-in-the-loop* studies that are based on traditional methods of qualitative analysis but employ NLP at some point in the process to increase speed and efficiency of analyses; and (d) *human-in-the-loop* studies that begin with NLP as expert but recruit a human researcher at some point in the analysis of data to augment the capabilities of NLP.

Human-as-expert studies assume that the human is the most accurate among human and artificial intelligence approaches to analyzing qualitative data. For example, when NLP is used to evaluate writing in high stakes tests such as the GRE and TOEFL [3], the educational testing service acknowledges that the NLP rater/evaluator "…doesn't have the ability to read so it can't evaluate essays the same way that human raters do." [3]. Instead, when NLP-generated scores are used to augment or complement human-generated scores, both measurement and reliability improve.

Thus, the human remains the expert on what constitutes a high quality vs. low quality essay, but NLP enhances the human's ability to be the expert.

In contrast, in lower stakes tests such as practice tests for the GRE and TOEFL, the NLP-based e-rater engine directly provides a score for a student's essay, considering features associated with grammar, usage, mechanics, style and organization, and development [3]. In this *NLP-as-expert* approach, NLP is solely responsible for scoring student essays and providing diagnostic feedback regarding writing quality. In many education research studies, the goal is also to make the NLP engine the expert at analyzing student and teacher data. Typically, these studies report the accuracy of the NLP approach compared to the human approach for a subset of the data or with the goal of relegating NLP to analysis of a much larger, future set of data. For example, Verleger et al. [9] reported accuracies of 60%-85% for NLP compared to traditional, human-based evaluation of student problem solving for open-ended engineering problems according to an 11-dimension grading rubric. The study was motivated by a need to automate the process of intelligently assigning peer reviewers to student teams in large classes. Similarly, another study [4] identified 70%-90% agreement between teachers' grades and grades generated by automated essay grading tools as a means to "…considerable reduction in essay grading costs." [4].

While in *NLP-as-expert* approaches, NLP is intended to be used as the sole assessor or classifier of text-based data, *NLP-in-the-loop* approaches retain the human element alongside the NLP algorithm. For example, Stratton et al, 2017 [15] used NLP to generate summaries of students' reflections collected across an entire semester. These summaries were then graded/assessed by human graders and used to quantitatively analyze how such reflection related to achievement goals. In another example of NLP-in-the-loop, Zhang et al. [22] used NLP to identify bias, unseen relationships, and missed coding opportunities among teachers' responses regarding questions related to the digital divide. The authors first used traditional methods of qualitative analysis to arrive at a set of thematic codes, then they used NLP techniques to cluster the survey responses and examined the semantic content captured by these techniques. They compared the themes resulting from the traditional approach to those arrived at through NLP to identify incongruities associated with errors and inconsistencies among human coders.

Our study focuses primarily on the fourth broad category of using NLP in education -- *human-in-the-loop*. This approach leverages the strengths of both NLP and human expertise in qualitative data analysis for overall improvements in results. The human expert provides a critical evaluation of the results generated by NLP, and the insights and judgments they provide are incorporated into the analysis to improve its accuracy and reliability. For example, Katz et al. [20] used NLP in the first stage of analyzing open-ended survey questions such as "What your [course] instructor have done differently in the online transition to help you learn?" to understand engineering student experiences during the COVID-19 pandemic. NLP was used to identify the major categories of words (i.e., topics) and then expert researchers converted those codes into themes for thematic analysis and interpretation of the qualitative data. Using this approach, the development of codes or themes was reduced from four to five hours per question using traditional approaches to under one hour for the NLP-assisted (human-in-the-loop) approach. Another study involving open-ended responses from teachers regarding their strategies to support student retention used a similar human-in-the-loop approach. Analysis began with topic modelling using NLP and the generation of dendograms which visually indicated relationships between topics. Both of these tools were then used by expert researchers to aggregate NLP generated codes into themes for thematic analysis of teachers' responses [21].

The success of *human-in-the-loop* approaches to more complex, multi-topic analysis of open-ended responses from students and teachers prompted their use in this study. Like many other studies, our work relies on thematic analysis to facilitate effective use of NLP in the processing of qualitative, text-based feedback from students. Unlike many previous studies, however, we apply NLP at a broader scope. Instead of asking students to reflect retroactively on their experiences in a particular course, we open the door wider, asking them to identify *anything at all* a TA could do to better support their learning.

Methods

This study is part of a larger, single-institution research project, which used a survey to investigate the connections between different forms of support (from faculty, TAs, and peers) and various dimensions of course-level engagement (including attention, participation, effort, and emotional engagement) in multiple learning contexts. The survey also included several short answer questions, one of which is analyzed in this study: "*What one action can your TAs at <this institution> take to best support you in your classes (please be as specific as possible)?"* Addressing this question using qualitative thematic analysis and NLP led to three research questions:

Educational Research Question (RQ1):
What do students most want from TAs to support their learning?
This question was analyzed using both traditional thematic analysis of student responses and using NLP-based thematic analysis. The comparison of NLP and traditional methods led to two additional research questions:

Methodology Research Question (RQ2):
How well does NLP Coding agree with Traditional (Human) Coding?

Methodology Research Question (RQ3):
Does NLP generate different conclusions than Traditional (Human) Coding?

Participants

This study recruited 1,454 undergraduates from primarily electrical and mechanical engineering majors between the fall of 2016 and the spring of 2021. Some students were taking courses in a traditional setting prior to the COVID-19 pandemic; other students were enrolled in emergency remote teaching (ERT) courses which were held remotely during the pandemic [23], [24]. The majority of students in this study (N = 1,071, 73.7%) were male and either Asian (N = 623, 42.8%) or White (N = 574, 39.5%). Most participants were US citizens or permanent residents (N = 1,229, 84.5%). Participant demographics are summarized in Table 1.

Procedures

This study was approved by the institutional review board (IRB) with the approval number STUDY00000378. The study recruited undergraduate students from various courses relevant to this research, but the researchers did not engage directly with the students. Participation in the study was voluntary, and students were informed that their survey responses would be kept confidential. Incentives in the form of extra credit were offered to students in several courses. The survey was administered electronically (online) in most courses but participants in one course completed paper copies of the survey.

Table 1. Demographics of study population (*N* = 1,454)

Demographic Variable	*N*	%	*N*	%	*N*	%
Gender	All Students		Traditional Setting		ERT Setting	
Male	1071	73.7%	318	73.4%	753	73.75%
Female	369	25.4%	112	25.8%	257	25.1%
Other	10	0.69%	3	0.69%	7	0.68%
Race	All Students		Traditional Setting		ERT Setting	
Asian	623	42.8%	153	35.3%	470	46.0%
White	574	39.5%	204	47.1%	370	36.2%
Black	31	2.13%	10	2.3%	21	2.05%
Other*	199	13.7%	61	14.0%	139	13.6%
U.S. Status	All Students		Traditional Setting		ERT Setting	
Domestic	1229	84.5%	376	86.8%	853	83.5%
International	217	14.9%	57	13.1%	160	15.6%

Percentages (of all respondents) may not add to 100% due to non-responses.
*Other: includes more than one (mixed) race, Native American, and Pacific Islander

Data Analysis

To preprocess the data, we used libraries including pandas, numpy, and sklearn in the Jupyter Notebook Python software. Figure 1 presents the approach for analyzing textual data obtained from short answer questions on TA support in the survey conducted for this study. Each step in the analysis is described in further detail, next.

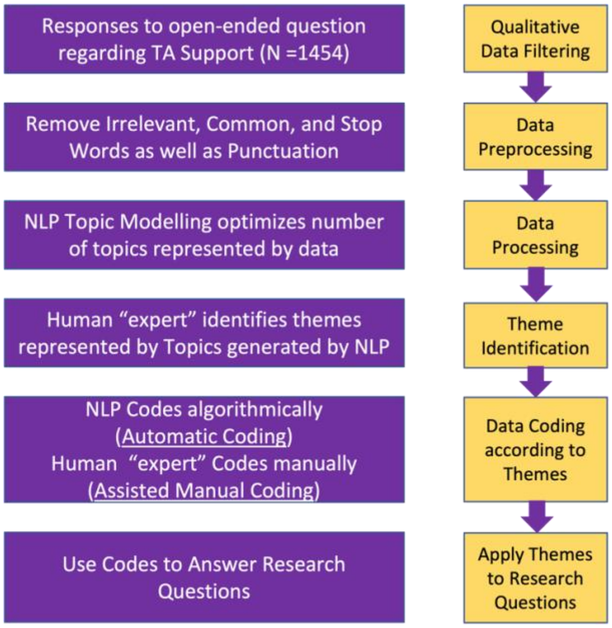
Responses to open-ended question regarding TA Support (N =1454)
Qualitative Data Filtering
Remove Irrelevant, Common, and Stop Words as well as Punctuation
Data Preprocessing
NLP Topic Modelling optimizes number of topics represented by data
Data Processing
Human "expert" identifies themes represented by Topics generated by NLP
Theme Identification
NLP Codes algorithmically (Automatic Coding)
Human "expert" Codes manually (Assisted Manual Coding)
Data Coding according to Themes
Use Codes to Answer Research Questions
Apply Themes to Research Questions

Figure 1: Data Analysis Approach

Qualitative Data Filtering: Of 1,586 original participants in this study, 132 either didn't respond to the TA support question, had no suggestions (e.g., “I don’t know”; “Nothing really”), or stated that they had no contact with their TAs. After these non-responses were deleted from the dataset to ensure the quality and relevance of the data for use in the topic modeling process, a total of 1,454 responses remained for our analysis. The resulting qualitative dataset is a solid corpus of data that is properly formatted and structured to support topic modeling for identifying patterns and generating insights into the research questions at hand.

Data Preprocessing: The word counts for the student responses to the short answer question were analyzed before and after preprocessing. Before preprocessing, the median word count was 14.0, and the maximum word count was 280. To further reduce the noise in the qualitative data, repeated words, punctuation, stop words (e.g., a, is, the), verbs deemed irrelevant by the domain expertise of the researchers (e.g., provide, make, get, hold, set), and words common across topics and responses (e.g., TAs) were eliminated using automated text cleaning techniques [25]. Data preprocessing also included converting all responses to lowercase. After preprocessing, the median word count was 6.0, and the maximum word count was 107. The data was then split into three sets: training (60%), validation (20%), and testing (20%). The resulting unstructured data sets were then transformed into structured data sets using the Count Vectorizer in the Python programming language [26] which converts text into vectors based on the frequency count of each word. Such vectorization is necessary to overcome the challenges posed by the unstructured and high-dimensional nature of the raw textual data.

Data Processing: The training data were used to initially optimize the number of topics emerging from the data (based on the words students used in their responses). The performance of the resulting topic model and hyperparameters were then optimized using the unseen data in the validation set. Once the model was optimized in training and validation, the accuracy and overall performance of the resulting topic model created by NLP was assessed using the unseen data in the testing set. Selecting the optimal number of topics to represent the training dataset was done using the elbow method. Student responses were then clustered into those topics using a form the Latent Dirchlet Algorithm or LDA (Figure 2).

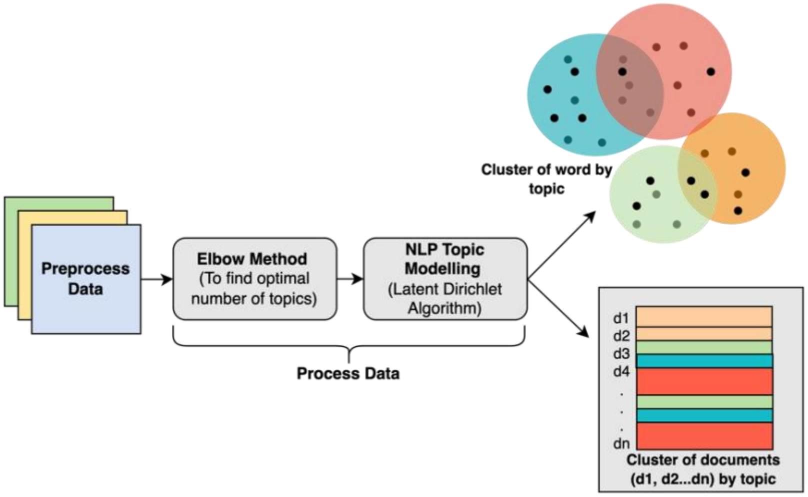
Cluster of word by topic
Preprocess Data
Elbow Method (To find optimal number of topics)
NLP Topic Modelling (Latent Dirichlet Algorithm)
Process Data
d1
d2
d3
d4
.
.
.
dn
Cluster of documents (d1, d2...dn) by topic

Figure 2: Data Processing

The elbow method involves applying *k*-means clustering to the pre-processed data for different values of *k*, starting from 1 and ending at a preset maximum value. In this study, *k* = 7 was chosen as a reasonable endpoint. For each *k*, the within-cluster sum of squared errors (SSE) is calculated and plotted in a scatter plot. The optimal number of topics is determined as the value of *k* at which the SSE value starts to decrease at an angle that resembles an elbow [27]. The optimal number of topics for the training dataset in this analysis was determined to be *k* = 4 (Figure 3). Once the optimal number of topics was determined, the data were processed using topic modeling, a machine learning technique that uses statistical models to extract hidden topics from textual data. Since the topics are not know in advance, this approach is considered unsupervised. In this study, one of the most popular and widely used topic modeling techniques, Latent Dirichlet Allocation (LDA) was applied to the data. LDA is a generative statistical model that represents a textual dataset as a combination of different topics, each defined by a unique set of words [28].

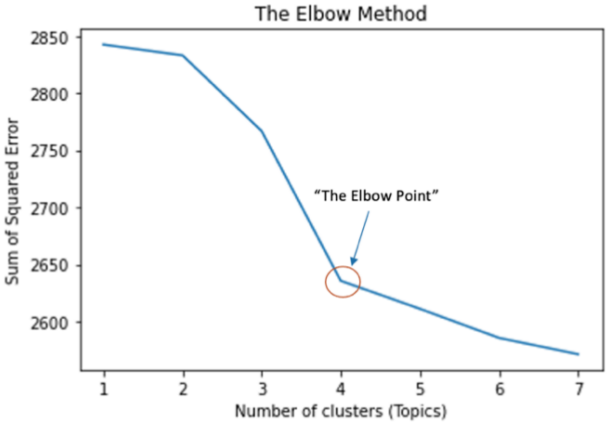

Figure 3: Determining the Optimal Number of Topics in NLP Topic Modelling

LDA can capture the underlying topics present in a corpus of text, even if different words with similar meanings are used to express those topics [28]. Mathematically, documents (i.e., student responses) are represented as a probability distribution over latent topics, while topics are represented as a probability distribution over words [29]. LDA generates both clusters of words and clusters of documents that are differentiated by their underlying topics.

Code Identification: The NLP topic modeling technique generates topics from the data by identifying the top word distributions based on their probability of occurrence in the text. These topics are represented as a collection of words with high probability of occurring together and are typically used to summarize the content of the text data. However, these topics are not necessarily grounded or contextualized in any theoretical framework. Thus, at this stage in the data analysis, a human researcher transforms the topics into themes that are appropriately grounded in teaching and learning theory. Theme identification involves first naming the topic (code) embodied by each collection of words assembled by LDA and then converting these topics/codes into themes. This process enhances the interpretability and usability of the topics generated by NLP-based topic modeling technique.

Data Coding: After codes were identified, all data (i.e., student responses) were labelled using automatic coding and manual coding. Automated coding allows the NLP algorithm to identify the code associated with each response using the topic model generated during *data processing*. This process requires no human intervention. Manual coding is a process where a human expert, in this case the researcher, uses these codes to label the students' responses. The human coder may apply more than one code to a student response, but the automated NLP-based technique only allows for one code per student response, Automated codes are then compared to manually assigned codes to determine for each topic/theme and classified as follows:

- True positives: indicate the number of times the automated coding and the corresponding manual coding agreed on the code assigned to a given response.
- True negatives: indicate the number of times the automated coding and the corresponding manual coding agreed that a given response did not belong to a particular code.
- False positives: indicate the number of times the automated coding assigned a code to a given response that was not assigned by the corresponding manual coding.
- False negatives: indicate the number of times the manual coding assigned a code to a given response that was not assigned by the corresponding automated coding.

These four metrics are used to answer our second research question and to assess the accuracy and usefulness of the NLP approach in processing and interpreting the data.

Apply codes to Research Questions: Finally, codes are converted to themes and the frequency of themes identified by manual coding are ranked and used to answer the first research question (RQ1) regarding the types of TA support that students prefer. To answer the last research question (RQ3), the process is repeated for the automated codes and conclusions generated by RQ1 and RQ3 compared to understand if, at the bottom line, the automated (NLP) approach reaches the same conclusions as the traditional (manual coding) approach.

Results

Four topics (codes) emerged from LDA topic modeling in our study sample: examples, office hours, questions and answers, and lab support. The most frequent words associated with each of these topics are summarized in Table 2. Two topics related to communications between students and TAs expressed in terms of out-of-class office hours (topic 2) and availability for questions and answers (topic 3). Both of these topics were identified as relevant to a theme of *interactions* between instructors and students which have been identified as one of the most if not the most influential contributor to satisfaction with college [30].

Table 2. Topics and Themes representing Student Responses regarding TA Support

Most Frequently Occurring Words associated with Each Topic			
Topic 1	***Topic 2***	***Topic 3***	***Topic 4***
problems, quiz, lecture, work, examples, homework, time, practice, clear, example	hours, office, available, time, times, hour, feedback, zoom, many, assignments	questions, answer, ask, discussion, emails, available, question, email, answering, online	lab, labs, extra, explain, things, time, online, especially, people, giving
Topic 1 Label: Examples	**Topic Label 2:** Office Hours	**Topic Label 3:** Questions and Answers	**Topic Label 4:** Lab Support
Theme 1: Engineering (Problem solving)	**Theme 2:** Interactions 1 (Office hours)	**Theme 3:** Interactions 2 (Q&A)	**Theme 4:** Active Learning (Experiential)

While the importance of interactions between students and instructors is a critical element of undergraduate education that is common to all fields and disciplines, the remaining two topics that emerged from topic modelling were more specific to engineering. Topic 1 emphasized student preferences for more problem-solving time and practice with TAs. This relates directly to the theme of *problem-solving* which is highlighted by the ABET (accreditation board for engineering and technology) student outcome #1:

> "an ability to identify, formulate, and solve complex engineering problems by applying principles of engineering, science, and mathematics." [31]

Topic 4 focused on students' need for TAs to provide effective laboratory support. Effective teaching through the lab experience is a theme related to *active learning* in engineering education. The experiential activities provided by engineering labs are critical to the transfer of learning from the classroom to the real-world.

Agreement (and disagreement) between automated NLP-based coding of student responses and manual (human) coding of these responses according to the topics and themes in Table 2 are presented in Table 3 for data seen during the development of the NLP model (i.e., the training set) and data not seen during model development (i.e., validation, and testing datasets). Two metrics, true positives, and true negatives, were used to describe agreement between the two coding methods and two metrics, false positives, and false negatives, were used to describe disagreement between them.

Table 3. Consistency between NLP and Human-Coding
(All values are given in Percentages of Responses)

<table>
<tr><th colspan="6">Training Data Results</th></tr>
<tr><th>Topics</th><th>True Positives</th><th>True Negatives</th><th>False Positives</th><th>False Negatives</th><th>Accuracy</th></tr>
<tr><td>Examples</td><td>29.7</td><td>49.6</td><td>9.53</td><td>11.1</td><td>79.3</td></tr>
<tr><td>Office Hours</td><td>13.5</td><td>71.1</td><td>5.97</td><td>9.42</td><td>84.6</td></tr>
<tr><td>Q&A</td><td>21.9</td><td>64.4</td><td>4.13</td><td>9.53</td><td>86.3</td></tr>
<tr><td>Lab Support</td><td>8.84</td><td>80.6</td><td>6.32</td><td>4.25</td><td>89.4</td></tr>
<tr><th colspan="6">Validation Data Results</th></tr>
<tr><th>Topics</th><th>True Positives</th><th>True Negatives</th><th>False Positives</th><th>False Negatives</th><th>Accuracy</th></tr>
<tr><td>Examples</td><td>26.8</td><td>52.6</td><td>8.59</td><td>12.0</td><td>79.4</td></tr>
<tr><td>Office Hours</td><td>17.5</td><td>68.7</td><td>7.22</td><td>6.53</td><td>86.2</td></tr>
<tr><td>Q&A</td><td>25.4</td><td>63.6</td><td>3.78</td><td>7.22</td><td>89.0</td></tr>
<tr><td>Lab Support</td><td>5.84</td><td>85.2</td><td>4.81</td><td>4.12</td><td>91.1</td></tr>
<tr><th colspan="6">Testing Data Results</th></tr>
<tr><th>Topics</th><th>True Positives</th><th>True Negatives</th><th>False Positives</th><th>False Negatives</th><th>Accuracy</th></tr>
<tr><td>Examples</td><td>28.5</td><td>52.6</td><td>10.7</td><td>8.25</td><td>81.1</td></tr>
<tr><td>Office Hours</td><td>15.8</td><td>70.1</td><td>4.12</td><td>9.97</td><td>85.9</td></tr>
<tr><td>Q&A</td><td>26.1</td><td>61.9</td><td>3.78</td><td>8.25</td><td>88.0</td></tr>
<tr><td>Lab Support</td><td>6.19</td><td>85.9</td><td>4.81</td><td>3.09</td><td>92.2</td></tr>
</table>

The NLP model found it most difficult to classify student responses regarding examples, with only 79.3%, 79.4%, and 81.1% accurately classified for training, validation, and testing datasets respectively. In contrast, student responses that related to lab support were accurately classified the most often with rates of 89.4%, 91.1%, and 92.2% for training, validation, and testing data respectively. Surprisingly, the accuracy rates (true positives + true negatives) for all four topics emerged as worst among data in the training set with only 79.3%, 84.6%, 86.3%, and 89.4% of student responses classified accurately for examples, office hours, Q&A, and lab support respectively.

Among student responses, both coding methods identified the topic of examples (corresponding to the theme of *problem solving*) as the most frequent form of instructional support that students desired from TAs. Across training, validation, and testing datasets, between 35.4% and 40.9% of student responses were coded into this category. Similarly, both (automatic and manual) coding methods identified the topic of lab support (corresponding to a theme of *active learning*) as the least frequent form of instructional support that students desired from TAs. Between 9.28% and 15.4% of students (dependent on dataset and coding method) indicated that lab support was where they needed TAs to support them most. Both themes of *interactions* (1 and 2) ranked in between themes of *problem solving* and *active learning* in terms of how frequent students expressed these themes in their preferences for TA support. A detailed summary of which student responses were categorized into what codes and themes is provided in Table 4.

Table 4. Automatic Coding (NLP) vs Manual Coding

<table>
<tr><th colspan="5">Training Data Results</th></tr>
<tr><td>Theme</td><td>Problem-Solving</td><td>Interactions 1</td><td>Interactions 2</td><td>Active Learning</td></tr>
<tr><td>Automatic Coding (NLP)</td><td>38.90%</td><td>25.90%</td><td>19.80%</td><td>15.40%</td></tr>
<tr><td>Assisted Manual Coding</td><td>40.90%</td><td>31.50%</td><td>23.00%</td><td>13.10%</td></tr>
<tr><th colspan="5">Validation Data Results</th></tr>
<tr><td>Theme</td><td>Problem-Solving</td><td>Interactions 1</td><td>Interactions 2</td><td>Active Learning</td></tr>
<tr><td>Automatic Coding (NLP)</td><td>35.40%</td><td>29.20%</td><td>24.70%</td><td>10.70%</td></tr>
<tr><td>Assisted Manual Coding</td><td>38.80%</td><td>32.70%</td><td>24.10%</td><td>9.97%</td></tr>
<tr><th colspan="5">Testing Data Results</th></tr>
<tr><td>Theme</td><td>Problem-Solving</td><td>Interactions 1</td><td>Interactions 2</td><td>Active Learning</td></tr>
<tr><td>Automatic Coding (NLP)</td><td>36.80%</td><td>25.80%</td><td>34.40%</td><td>9.28%</td></tr>
<tr><td>Assisted Manual Coding</td><td>39.20%</td><td>19.90%</td><td>29.90%</td><td>11.00%</td></tr>
<tr><th colspan="5">Overall Data Set</th></tr>
<tr><td>Theme</td><td>Problem-Solving</td><td>Interactions 1</td><td>Interactions 2</td><td>Active Learning</td></tr>
<tr><td>Automatic Coding (NLP)</td><td>37.78%</td><td>26.54%</td><td>23.70%</td><td>13.24%</td></tr>
<tr><td>Assisted Manual Coding</td><td>40.14%</td><td>29.42%</td><td>24.60%</td><td>12.05%</td></tr>
</table>

Overall, the results indicate that the automated coding methods based on NLP topic models returned the same conclusions as did traditional, manual coding methods. And further, the model's performance on the test and validation data is consistent with the results from the training data indicating that the topic model is neither overfitted or too general.

Discussion

This research aimed to examine the use of Natural Language Processing (NLP) in the analysis of student opinions and feedback on instructional support and more specifically, what students desired from TAs to support their learning. A human-in-the-loop approach was used, which leveraged both NLP and human expertise in qualitative data analysis. The study yielded definitive and promising answers to all three research questions.

Educational Research Question (RQ1)

What do students most want from TAs to support their learning?

Both the NLP topic modeling algorithm and traditional thematic analysis of the textual data in this study indicated that engineering students in the context of this single institution primarily want additional support with their coursework from TAs through practice in *problem-solving*. Students expressed this theme in the context of quizzes, homeworks, and other assignments; they wanted TAs to provide more problem-solving practice across the board. The prevalence of this theme in student preferences is consistent with student learning outcomes expected for accreditation in engineering programs but is not a generalized priority in higher education.

NLP also discovered that interactions between students and instructors (including TAs) are a high priority for the engineering students in our study; this is consistent with all of higher education which has demonstrated ample evidence that interactions with faculty/instructors have a profound impact on college satisfaction [30]. Students in engineering are no exception to the importance of instructor interactions and they expressed this in their preferences for TA support with moderate frequency in terms of both office hours (topic 2) and question and answer opportunities (topic 3). Somewhat surprisingly, though, a lower proportion of students prioritized lab support in their expectations of TAs. Considering the importance of the laboratory in learning transfer and experiential learning in education, we expected that more student preferences would reflect greater support in this area of their education. Whether this result is a function of the population studied or is more generalizable remains a question to be explored in future work. Nevertheless, these findings highlight the importance of TAs in the learning process of students, and the need for institutions to invest in TA training programs and support services. Providing TAs with the resources and support they need to effectively meet the diverse needs of students can help improve students' academic outcomes and support their overall success. This and similar studies provide important evidence regarding what the priority areas should be for enhancing TA support for engineering students.

Methodology Research Question (RQ2)

In this case study, how well does NLP Coding agree with Traditional (Human) Coding?

The study evaluated the validity of using NLP to classify/code the frequency of student responses with regard to their preferences for instructional support. Classification accuracy ranged from 79% to 92%, well above what would have been expected by chance. Thus, the results of this study add to the growing body of evidence supporting NLP as an effective application of machine learning to qualitative data analysis in education research.

Methodology Research Question (RQ3)

In this case study, does NLP generate a different conclusion than Traditional (Human) Coding?

Overall, our results indicate the same conclusions from NLP (automated) coding of qualitative data compared to traditional (manual) coding (Table 4) in terms of ranking themes in student preferences for TA support. This result further supports the use of NLP in engineering education research.

Limitations

This study is limited by potential biases introduced by human annotators during pre-processing and its narrow focus on a single US research institution. The results from the first research question (RQ1) may not apply to other institutions. The racial composition was also not representative of the overall engineering enrollment in the US, with Asian American students overrepresented (43.6% vs. 15.1% nationally) [32] and Black students underrepresented (2.17% vs. 4.80% nationally) [32]. The courses studied were limited to two engineering disciplines with low representation of women students, but the percentage of women in the sample (25.4%) slightly exceeded the national average (23.6%) for bachelor's degrees in engineering [32]. The themes of TA support are likely to be present in other engineering student populations, but their priority may differ. The use of NLP in thematic analysis in this study, combined with human input, may or may not produce similar results if applied to a different student population. However, the third research question (RQ3) found that the conclusions from both NLP and human thematic analysis were the same, indicating that NLP can effectively and accurately reduce human resources needed for qualitative research and data analysis.

Implications

As our research team and other teams invested in engineering education research continue to explore and optimize the role that NLP can play in qualitative data analyses, the potential for initiating qualitative research and integrating it into existing quantitative research designs to gain deeper and broader insight into important research questions is significant. NLP not only offers a way to reduce the human resources required for qualitative data analysis, but it can also be used to reveal bias in "expert" human coding. Although NLP approaches are intrinsically biased, they are not typically not biased in the same way as human coders.

In this study, for instance, topic modeling was used to understand student responses to questions regarding TA support. Inherently, this biased the themes and conclusions of the study to the words that students used in their responses. In contrast, sentiment analysis would influence these themes and conclusions based on the emotions reflected in student responses. But, neither topic modeling nor sentiment analysis would, in principle, be significantly biased toward language usage styles favored by specific races, genders, socioeconomic groups, etc. Human coders, on the other hand, are vulnerable to such biases in the use of language as well as biases introduced by the theoretical perspective, they employ to frame research designs.

Conclusions

This paper has presented a qualitative data analysis case study using NLP for engineering education research. The study compares qualitative thematic analysis methods with human-in-the-loop NLP methods. NLP-based methods identified four main topics/codes in student responses regarding preferred TA support: examples, office hours, questions and answers, and laboratory support, which an "expert" human coder converted into three themes: problem-solving, interactions (1 and 2), and active learning. It is important to note that these findings are contextual and situational and may not necessarily generalize to all engineering students. We found that traditional and NLP-based methods agree (accuracy) at rates between 79.3% and 92.2% across the four codes (and three themes) associated with this analysis. Most importantly, traditional thematic analysis and NLP-based (human-in-the-loop) analysis reached the same conclusions about engineering student TA support. The most frequent theme in the responses of students was for TAs to provide them with high quality *problem-solving* support. Second and third ranked topics were related to a theme of *interactions* between students and TAs; many students ranked it their top priority for TAs to offer opportunities to be available for questions, answers, and discussion. And, although reported less frequently, some students thought it most important for TAs to fully support their laboratory or *active/experiential* learning in their engineering courses.

This start-to-finish analysis of text-based data from open-ended survey questions added to existing evidence that NLP can be a powerful tool for qualitative analysis and research in education. Future research should optimize NLP use in qualitative analyses and demonstrate its efficacy in further expanding qualitative research capacity in engineering education research. Future research will also explore code and theme frequency by gender, race, and ethnicity and also explore error rates among those different groups.

References

[1] E.D. Liddy, "Natural Language Processing," in *Encyclopedia of Library and Information Science*, 2nd Ed., NY, Marcel Decker, Inc., 2001.

[2] S. Tenny, J. M. Brannan, G. D. Brannan, *Qualitative Study*. Treasure Island, FL: StatPearls Publishing, 2022. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470395/

[3] "About e-rater", *Educational Testing Service (ETS).* [online]. Available: https://www.ets.org/erater/about.html

[4] A. Shehab, M. Elhoseny, and A. E. Hassanien, "A hybrid scheme for Automated Essay Grading based on LVQ and NLP techniques," in *2016 12th International Computer Engineering Conference (ICENCO), Cairo, Egypt, December 28-29, 2016*, doi: 10.1109/icenco.2016.7856447.

[5] L. K. Allen., and D. S. McNamara, "You Are Your Words: Modeling Students' Vocabulary Knowledge with Natural Language Processing Tools," in *International Educational Data Mining Society, Paper presented at the International Conference on Educational Data Mining (EDM), Madrid, Spain, Jun 26-29, 2015*.

[6] S. Ghosh, "*Online Automated Essay Grading System as a Web Based Learning (WBL) tool in Engineering Education,"* in Web-based Engineering Education: Critical Design and Effective Tools, IGI Global, 2010, pp. 53-62, doi: 10.4018/978-1-61520-659-9.ch005.

[7] V. M. Holland and J. D. Kaplan, "Natural language processing techniques in computer assisted language learning: Status and instructional issues," *Instructional Science*, vol. 23, no. 5-6, pp. 351-380, Nov. 1995, doi: 10.1007/bf00896878.

[8] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, "Classification of open-ended responses to a research-based assessment using natural language processing," *Physical Review Physics Education Research*, vol. 18, no. 1, June 2022, doi: 10.1103/physrevphyseducres.18.010141.

[9] M. A. Verleger, "Using Natural Language Processing Tools to Classify Student Responses to Open-Ended Engineering Problems in Large Classes," in *2014 ASEE Annual Conference & Exposition Proceedings*, *Indianapolis, Indiana, June 15- 18, 2014*, pp. 24.1338.1 - 24.1338.15, doi: 10.18260/1-2--23271.

[10] A. Agogino, S. Song, and J. Hey, "Triangulation of Indicators of Successful Student Design Teams," *International Journal of Engineering Education*, vol. 22, no. 3, pp. 617-625, 2007.

[11] S. A. Crossley, D. R. 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?," *The Journal of Writing Analytics*, vol. 1, no. 1, pp. 48–81, 2017, doi: 10.37514/jwa-j.2017.1.1.04.

[12] S. Lim, C. S. Tucker, K. Jablokow, and B. Pursel, "Quantifying the Mismatch Between Course Content and Students' Dialogue in Online Learning Environments," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Cleveland, Ohio, August 6-9, 2017,* vol. 58158, p. V003T04A016.

[13] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, "Forecasting student achievement in MOOCs with natural language processing," *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16, Edinburgh United Kingdom April 25 - 29, 2016*, pp.383-387, doi: 10.1145/2883851.2883932.

[14] 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," in *International Conference on Educational Data Mining (EDM), Raleigh, NC, Jul 16-20, 2018.*

[15] D. H. Stratton, S. Anwar, and M. Menekse, "How Do Engineering Students' Achievement Goals Relate to their Reflection Behaviors and Learning Outcomes?," in *Proceedings of the 2017 ASEE Annual Conference & Exposition, Columbus, Ohio, June 24-28, 2017*, pp. 1-10, doi: 10.18260/1-2—28444.

[16] S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara, and R. S. Baker, "Combining click-stream data with NLP tools to better understand MOOC completion," *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, - LAK '16, Edinburgh United Kingdom, April 25 - 29, 2016,* pp. 167-171, doi: 10.1145/2883851.2883931.

[17] M. Soledad, J. Grohs, S. Bhaduri, J. Doggett, J. Williams, and S. Culver, "Leveraging instit utional data to understand student perceptions of teaching in large engineering classes," in *2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA, October 18-21, 2017,* pp. 1-5, doi: 10.1109/fie.2017.8190608.

[18] 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), Muscat, Oman, March 15-16, 2016,* pp. 1–5, doi: 10.1109/ICBDSC.2016.7460390.

[19] A. Satyanarayana, K. Goodlad, J. Sears, P. Kreniske, M. 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," in *2019 ASEE Annual Conference & Exposition Proceedings, Tampa, Florida, June15-October 19, 2019,* doi: 10.18260/1-2--31917.

[20] A. Katz, M. Norris, A. M. Alsharif, M. D. Klopfer, D. B. Knight, and J. R. Grohs, "Using Natural Language Processing to Facilitate Student Feedback Analysis," in *2021 ASEE Virtual Annual Conference Content Access, July 26-29, 2021.* [online]. Available: https://peer.asee.org/using-natural-language-processing-to-facilitate-student-feedback-analysis

[21] D. Buenano-Fernandez, M. Gonzalez, D. Gil, and S. Lujan-Mora, "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach," *IEEE Access*, vol. 8, pp. 35318–35330, 2020, doi: 10.1109/access.2020.2974983.

[22] T. Zhang, M. Moody, J. P. Nelon, D. M. Boyer, D. H. Smith, and R. D. Visser, “Using Natural Language Processing to Accelerate Deep Analysis of Open-Ended Survey Data,” presented at *2019 SoutheastCon, Huntsville, AL, USA*, Apr. 2019, doi: 10.1109/southeastcon42311.2019.9020561.

[23] C. Hodges, S. Moore, B. Lockee, T. Trust, and A. Bond, “The difference between emergency remote teaching and online learning,” *Educause review*, vol. 27, pp. 1-12, 2020. [online]. Available: https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning.

[24] T. A. Ghebreyesus, "WHO Director-General's opening remarks at the media briefing on COVID-19," in *WHO Director-General's Media Briefing on COVID-19*, 11 March 2020. [online]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. [Accessed: 2-Feb-2023].

[25] M. Makrehchi and M. S. Kamel, “Extracting domain-specific stopwords for text classifiers,” *Intelligent Data Analysis*, vol. 21, no. 1, pp. 39–62, Jan. 2017, doi: 10.3233/ida-150390.

[26] Sklearn.org. "CountVectorizer." sklearn.feature_extraction.text, *scikit-learn.org*, 2018. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html. [Accessed 2-Feb-2023].

[27] H. Humaira and R. Rasyidah, “Determining the Appropriate Cluster Number Using Elbow Method for K-Means Algorithm,” in *Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018*, 24-25 January 2018, Padang, Indonesia, 2020, doi: 10.4108/eai.24-1-2018.2292388.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[29] F. Gurcan and N. E. Cagiltay, “Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling,” *IEEE Access*, vol. 7, pp. 82541–82552, 2019, doi: 10.1109/access.2019.2924075.

[30] A. W. Astin, "Student involvement: A developmental theory for higher education," *Journal of College Student Personnel*, vol. 25, no. 4, pp. 297-308, 1984.

[31] Criteria for Accrediting Engineering Programs, 2022-2023, *Accreditation Board for Engineering and Technology (ABET)*. [online]. Available: https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2022-2023/. [Accessed 6-Feb-2023]

[32] “Engineering and Engineering Technology by the Numbers, 2021,” *American Society for Engineering Education (ASEE)*. [online]. Available: https://ira.asee.org/wp-content/uploads/2022/09/Engineering-and-Engineering-Technology-by-the-Numbers-2021.pdf