A Comparative Analysis of Natural Language Processing Techniques for Analyzing Student Feedback about TA Support

Neha Kardam, University of Washington

Neha Kardam is a PhD candidate in Electrical and Computer Engineering at the University of Washington, Seattle. She is an interdisciplinary researcher with experience in statistics, predictive analytics, mixed methods research, and machine learning techniques in data-driven research.

Dr. Denise Wilson, University of Washington

Denise Wilson is a professor and associate chair of diversity, equity, and inclusion in electrical and computer engineering at the University of Washington, Seattle. Her research interests in engineering education focus on the role of self-efficacy, belonging, and instructional support on engagement and motivation in the classroom while her engineering workplace research focuses on the role of relatedness, autonomy, and competence needs on persistence and fulfillment.

Sep Makhsous, University of Washington

A Comparative Analysis of Natural Language Processing Techniques for Analyzing Student Feedback about TA Support

Abstract

This paper advances the exploration of Natural Language Processing (NLP) for automated coding and analysis of short-answer, text-based data collected from student feedback. Specifically, it assesses student preferences for Teaching Assistant (TA) support in engineering courses at a large public research university. This work complements existing research with an in-depth comparative analysis of NLP approaches to examining qualitative data within the realm of engineering education, utilizing survey data (training set = 1359, test set = 341) collected from 2017 to 2022. The challenges and intricacies of multiple types of classification errors emerging from five NLP methods are highlighted: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), BERTopic, Latent Semantic Analysis (LSA), and Principal Component Analysis (PCA). These results are compared with results from traditional thematic analysis conducted by a domain expert to assess their efficacy. Two principal findings emerged for TA teaching practice and for the use of NLP in education research. Firstly, the conclusions derived from each coding technique are consistent, demonstrating that students want, in order of priority, extensive TA-student interactions, problem-solving support, and experiential/laboratory learning support from TAs. Secondly, the research offers insights into the effectiveness of NMF and PCA in processing educational survey data. A comparison of NMF and PCA topic models, based on accuracy, precision, recall, and F1 scores, reveals that while PCA outperforms NMF in terms of precision (identifying truly relevant topics), NMF excels in recall (capturing a broader range of student responses).

Introduction

The landscape of educational data collection is rapidly evolving, with significant increases in student enrollments and class sizes leading to an unprecedented growth in textual data from academic sources, such as assignments, assessments, and student feedback instruments [1] - [3]. This proliferation of textual data presents a critical challenge: manual analysis methods are increasingly untenable due to their time-intensive nature, highlighting the necessity for automation in the assessment process, whether in whole or in part [4]. In response to this need, a significant body of recent research has focused on the use of Natural Language to assess student work in the form of short answers, essays, or other formats. Far less research has focused on the automated analysis of student feedback collected from surveys and similar instruments. Responses to short answer questions in educational surveys can differ from text generated by students in assessment of their performance. Specifically, students are more likely to veer off topic or introduce ambiguity in their responses when they know that their answers are not being graded or otherwise assessed in a way that affects their academic performance and record.

To expand on the potential of NLP to automate the coding and analysis of student feedback in educational research, this study focuses on methods that involve both machine learning (NLP) and traditional, domain expert intervention. It applies these methods to a dataset that is large ($n > 1500$) compared to many qualitative research and analysis studies. Utilizing this comprehensive dataset of student responses, our study not only investigates the feasibility of NLP for student feedback and educational research data analysis but also explores the potential to enhance the objectivity and efficiency of coding processes. By comparing the outcomes of

NLP techniques with traditional coding methods, this research contributes to the ongoing discourse on the integration of automated technologies in educational settings. It aims to provide evidence-based insights into the comparative advantages and limitations of NLP, thereby informing future applications of these technologies in educational assessment and research.

Background

NLP is an interdisciplinary field encompassing machine translation, text processing, and artificial intelligence. It has emerged as a powerful tool for automating the evaluation of textual data in educational settings [5]. Research in the use of NLP in education has delved into the comparative analysis of NLP coding techniques with traditional manual coding methods, aiming to assess the reliability, validity, and efficiency of automated approaches [6] - [10]. These comparative studies have sought to identify the strengths and limitations of NLP technologies in capturing the nuances of student language expression, as well as their potential to replace or complement human expertise in data analysis processes [11].

A majority of this research has focused on using NLP to analyzing text-based data to assess student performance. Efforts to analyze student essays have been demonstrated in higher education in the fields of English language learning [12], psychology [13], and even in STEM fields including physics [14]. Essays are not as common an instrument in STEM fields as in liberal arts disciplines because assessment often emphasizes problem solving and the grasp of specific concepts. Thus, short answer question analysis is more common in STEM fields [15] and NLP has been demonstrated for assessing learning via these types of questions in such disciplines as computer science [16] and engineering [8].

Compared to the body of research exploring the use of NLP to assess student learning through essays, short answer questions, and similar instruments, existing research focusing on NLP for analyzing student feedback remains relatively sparse. Within this limited space, Katz et al. [8] used NLP techniques to reduce the dimensionality of a large textual dataset collected from student responses regarding the transition from traditional teaching to COVID-19 (pandemic) era teaching. NLP was used to reduce the data to a manageable set of clusters that could be subsequently analyzed thematically by a domain expert, thereby improving the speed and accuracy of the qualitative data analysis process. Similarly, Buenano-Fernandez et al. [11] utilized topic modeling methods within NLP to analyze self-assessment responses from teachers. They created a network to visualize how these extracted topics interrelated. Other, more limited studies have explored sentiment analysis to gauge student satisfaction or used NLP for basic classification tasks within student feedback data.

While most research focuses on NLP for automated scoring of student work, a study by Kerkhof [10] analyzing open-ended question scoring techniques highlights the potential of NLP for broader analysis of student feedback in educational research. This review highlights three key areas for applying NLP to open-ended questions: data pre-processing (cleaning and normalization), processing data through feature extraction (including semantic similarity measures using or knowledge-based approaches), and finally, clustering data to group similar responses based on the extracted features [17]. This breakdown showcases the potential of NLP for tasks beyond just scoring student work, but for also for analyzing and understanding what students are saying and feeling in a deeper way. Some studies have also used techniques like

Latent Dirichlet Allocation (LDA) for topic modeling to categorize key themes in student comments [7]. Similarly, researchers have leveraged Non-Negative Matrix Factorization (NMF) to extract meaningful insights from course evaluation feedback [8]. Additionally, Principal Component Analysis (PCA) has been used to categorize and understand student perceptions of various educational aspects [8].

In the NLP research, a comprehensive analysis comparing the effectiveness of various NLP techniques for analyzing student feedback remains largely absent. This paper seeks to provide such a comparative study and to complement the existing research by adding further empirical evidence that NLP is indeed a valuable tool for education. Via comparative analysis of different NLP techniques, this research aims to identify the approaches best suited to extract meaningful insights from the rich and nuanced data that emerges from student feedback in engineering education.

Like much NLP research, previous studies have focused on metrics such as accuracy, precision, and recall to capture the goodness of different approaches. While these metrics are important and meaningful to the NLP and machine learning community, they are not necessarily accessible or important to educational researchers. To bridge this gap, this study also adds to existing work by comparing the performance of multiple NLP approaches in term of these traditional technical metrics and also in terms of the bottom-line conclusions derived from each approach.

Research Questions

A comparative analysis approach to analyzing student feedback regarding their preferences for TA support yielded the following two research questions:

Research Question #1 (RQ1):
How accurate are different NLP techniques in interpreting student feedback?
Various NLP techniques, including Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), BERTopic, Latent Semantic Analysis (LSA), and Principal Component Analysis (PCA) are compared in terms of the accuracy with which they interpret qualitative data from student feedback. The comparative analysis aims to identify the strengths and weaknesses of these NLP techniques compared to each other and compared to traditional qualitative data analysis techniques (specifically thematic analysis by a domain expert).

With the hypothesis that NLP alone is not sufficient for textual data which is broad in scope and often ambiguous, this study also addresses the following question regarding hybrid approaches that use both traditional (human-based) and automated approaches to qualitative data analysis:

Research Question #2 (RQ2):
Can NLP replace domain expert coding in processing student feedback?
This question compares the performance of hybrid NMF and PCA analysis with traditional, fully manual domain expert coding using Cohen's kappa to understand the potential for hybrid methods to be used more extensively by education researchers.

Methods

This paper is part of a comprehensive research project conducted within a single institution across multiple academic years. The overall goal of the research is to explore the relationship between different types of learning support (provided by faculty, teaching assistants (TAs), and peers) and various aspects of engagement at the course level, encompassing both behavioral and emotional dimensions, across diverse learning environments including traditional and remote settings [7]. The survey used to support this research incorporated several short-answer questions to gain deeper insights into instructional support strategies most effective for engineering students. Student participants were asked to articulate their preferences regarding the ways in which peers, faculty, and TAs could offer support. Notably, one of the pivotal, short-answer questions guiding this study was: *"What specific action could TAs at <this institution> take to offer you the most effective support in your classes?"*

Participants

This study recruited 1,855 undergraduates, spanning lower division (200 and 300 level) to upper division (300 and 400 level) undergraduate courses between the fall of 2017 to the spring of 2022. The participants' experiences varied depending on when they took the course. Some students participated in traditional, in-person classes during pre-pandemic semesters (2017, 2018) or in a post-pandemic return to in-person learning (2022). Others experienced courses during the COVID-19 pandemic (2020 and 2021) when traditional teaching transitioned to emergency remote instruction [20]. 38.7% of students responding to the survey completed it while enrolled in traditional learning settings while 61.3% completed it during remoting learning (Emergency Remote Teaching or ERT) during the peak of the COVID-19 pandemic.

The gender composition of the student population in this study was representative of undergraduates enrolled in engineering programs in the U.S. The majority of students (n = 1376, 74.1%) in this study were male, compared to national representation where 74.1% of students in engineering are male [21]. Some races and ethnicities were not as well represented in this study compared to national data. Black (or African American) students were significantly underrepresented in this study, making up only 2.20% of the overall study population compared to 5.40% representation nationally as were Latino/a student's (3.69% in our study, compared to 15.8% of undergraduate engineers reporting as Hispanic at a national level) [21]. In contrast, Asian American students were highly over-represented at 19.2% of the study population compared to 16.1% nationally in engineering. discipline [21]. International or foreign students were also overrepresented in this study and consisted of 17.7% of the participant population compared to 7.9% of engineering students nationally [21].

Procedures

This study was approved by the institutional review board (IRB) with the internal approval number STUDY00000378. The study recruited undergraduate students from 3 courses in mechanical engineering and 18 courses in electrical and computer engineering resulting in a study population that included students majoring in these two disciplines as well as students in other engineering and physical science disciplines (e.g., physics). 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* = 1,855)

Demographic Variable	*N*	%	*N*	%	*N*	%
	All Students		Traditional Setting		ERT Setting	
Total	1842	100%	718	38.7%	1137	61.29%
Gender						
Male	1376	74.1%	532	74.0%	844	73.2%
Female	452	24.4%	176	24.5%	276	24.2%
Other	14	0.75%	5	0.69%	9	0.79%
Race	All Students		Traditional Setting		ERT Setting	
Asian American	358	19.2%	6	0.83%	352	31.0%
Asian International	218	11.7%	66	9.20%	152	13.3%
Black	40	2.20%	17	2.36%	23	2.02%
Latino/a	67	3.61%	28	3.90%	39	3.43%
White	699	37.7%	289	40.2%	410	36.0%
Mixed Asian/White	101	5.44%	35	4.87%	66	5.60%
Other*	372	20.0%	277	38.5%	95	8.35%
U.S. Status	All Students		Traditional Setting		ERT Setting	
Domestic	1526	82.2%	609	84.8%	917	80.6%
International	329	17.7%	109	15.1%	220	19.3%

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

Data Analysis

The survey data were first cleaned to remove responses from students who did not consent to the research. Demographics were then aggregated to complete the gender, racial, ethnicity, and learning context breakdowns provided in Table 1. Survey responses from students who did not complete the TA support question (e.g., responded "I don't know"; "Nothing really" or stated that they had no contact with their TAs) were also deleted [7]. This resulted in 1700 total responses for subsequent analysis. The data were then pre-processed by converting all responses to lowercase and removing stopwords, punctuation, and non-ASCII values. The pre-processed data were randomly divided into an 80% training set comprising 1359 instances and a 20% test set comprising 341 instances. To convert the unstructured data into structured data, the countvectorizer from the Sklearn library using Python software was employed, transforming the text into vectors based on the frequency count of each word [22].

To manage the high dimensionality of the raw textual data, the count vectorizer's vocabulary was limited to words that repeated in less than 80% of the responses but could also be found in at least two responses. The resulting structured data were fed into five topic modeling techniques: latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), bidirectional encoder representations from transformers (BERTopic), latent semantic analysis, and principal component analysis (PCA). Each technique offers a unique approach to uncovering meaningful insights into the textual data.

Latent Dirichlet Allocation (LDA) represents the dataset into topics based on word distributions, which aids in understanding key words and how they relate to the topics [23]. In contrast, non-negative matrix factorization (NMF) performs unsupervised clustering and dimensionality reduction, often using TF-IDF (Term Frequency-Inverse Document Frequency), a metric that assesses word importance by considering both its frequency within a single document (such as a survey response) and its rarity across all documents [24]. BERTopic, on the other hand, leverages BERT (Bidirectional Encoder Representations from Transformers) model for topic modeling to generate topic clusters that consider both local and global contexts, providing meaningful topic insights [25]. Latent Semantic Analysis (LSA) uses a word-document matrix to capture word frequencies, applying mathematical transformations to preserve essential word-document relationships in a lower-dimensional space [26]. Finally, Principal Component Analysis (PCA) simplifies high-dimensional datasets by analyzing covariance matrices, revealing patterns and structures that enhance understanding and visualization [27].

Each method was used to classify data into 3-7 topics. The optimal number of topics was determined based on the lowest perplexity score, signifying the model's improved generalization performance [28], [29]. This metric, widely used in probabilistic or text modeling, measures the model's predictive power and its ability to handle unseen data. A lower perplexity score indicates that the model has a higher certainty in its predictions [30]. Topics identified by NLP algorithms were then reduced to a smaller number of themes by a domain expert in engineering education by reviewing the top words that emerged in each topic. The domain expert then analyzed each student response and assigned it to one or more themes or alternatively, identified the response as ambiguous (i.e., not applicable to any theme). These results were then compared to the different NLP modelling approaches using the optimal number of topics determined during the topic-modelling phase of analysis. All NLP techniques were hybrid, in that they were not fully automated but relied on a domain expert to aggregate topics into appropriate themes and to identify ambiguous responses for deletion or consideration in the subsequent comparative analysis.

For comparative analysis, the results of data analysis were evaluated by (a) comparing NLP modelling techniques to traditional, domain expert analysis using technical performance metrics that are widely used in the NLP community as well as overall conclusions regarding the meaning and message of the results; and (b) using Cohen’s Kappa to analyze interrater reliability between themes identified by top NLP modelling techniques emerged and those assigned by the domain expert [31].

The following performance metrics formed the basis of comparison among the five NLP techniques explored in this study:

- True Positive (TP): The number of correctly identified positive observations. In the context of this study, TP represents the instances where the model correctly identifies students' responses related to TA support experience [32],[33].
- True Negative (TN): The number of correctly identified negative observations. In the context of this study, TN represents instances where the model correctly excludes responses not related to TA support experience [32],[33].

- False Positive (FP): The number of incorrectly identified positive observations. FP represents the instances where the model incorrectly identifies a response as related to a particular TA support theme [32],[33].
- False Negative (FN): The number of incorrectly identified negative observations. FN represents the instances where the model fails to identify responses related to a particular TA support theme [32],[33].
- Precision: Precision is the ratio of true positive predictions to the total predicted positives. Higher precision means fewer false positives, indicating the accuracy of the model in correctly identifying students' responses related to TA support experience [32],[33].
- Recall: Also known as sensitivity, recall is the ratio of true positive predictions to the total actual positives. Higher recall means fewer false negatives, ensuring that all instances of TA support are identified [32],[33].
- F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balance between precision and recall, and is especially useful when there is an uneven class distribution [32],[33].

In addition to these performance metrics, the following metrics were also examined to compare results from the educational researcher's perspective:

- Ranking: Themes identified by each NLP technique were ranked by frequency and compared to each other and to domain expert coding to understand whether NLP approaches reached similar conclusions to traditional methods of analyzing the data.
- Cohen's Kappa: Cohen's Kappa is a quantitative performance metric for classification models that assesses agreement between two raters. In this study, it was used to assess the agreement between the domain expert and the most promising NLP modelling techniques that emerged from comparative analysis. Values of Cohen's Kappa above 0.75 were considered excellent agreement, between 0.4 and 0.75 fair to good agreement and less than 0.4 were considered poor agreement. These ranges are consistent with current conventions for assessing interrater reliability [31]. Cohen's Kappa was calculated for each theme in the data, using 2X2 contingency tables that evaluated how well a particular theme identified by the domain expert agreed with the theme assigned by top NLP modelling techniques classification models.

Results

In our study sample, initial topic modeling revealed the emergence of four topics (also referred to as codes). Table 2 displays the most frequently appearing words linked with each of these four topics. Topic 1 reflected students' desire for greater practice with solving problems associated with engineering content including but not limited to additional examples, practice quizzes, and detailed homework solutions. All responses associated with Topic 1 were subsequently assigned to a theme of "problem solving". Topic 4 referred to student concerns about sufficient TA assistance in explaining and completing laboratories and other experiential or active learning activities; this topic was placed into a theme labelled "experiential learning." Both Topic 2 and Topic 3 indicated students' preferences for increased interaction with TAs, including extended office hours, online (Zoom) meetings, email correspondence, and other forms of question-and-answer engagement. These codes were combined under a single theme of "TA-student interactions". These three themes are all an essential aspect of engineering education as

highlighted by the accreditation board for engineering and technology (ABET) student outcomes [34].

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
Theme 1: Problem solving	**Theme 2:** TA-Student Interactions		**Theme 3:** Experiential Learning

The analysis of the data, as shown in Table 3, reveals consistent findings across all five NLP techniques and manual domain expert coding regarding the importance of each theme of TA support to students. The majority of students, in both the training and testing datasets, expressed a preference for increased interactions with TAs, while the fewest number of students showed support for experiential learning, which includes laboratories and other active learning activities. This trend was consistent across all NLP coding techniques, with problem-solving support ranking between TA-student interactions and experiential learning support in terms of student preference.

The NLP coding techniques that showed the highest agreement with domain expert coding on average were BERTopic for the training data and PCA for the testing data. Additionally, the third highest agreement was observed for the NMF (Non-negative Matrix Factorization)

Table 3. Results Summary for NLP Coding vs Domain Expert Coding

Training Data Results (N=1359)						
Theme	**Domain Expert Coding**	**NLP Coding (NLP)**				
		LDA	***BERTopic***	***NMF***	***LSA***	***PCA***
Problem Solving	30.2%	26.9%	48.9%	33.3%	27.3%	29.2%
TA-Student Interactions	54.2%	68.2%	52.2%	50.3%	77.7%	68.3%
Experiential Learning	10.2%	27.7%	10.5%	21.0%	14.9%	17.3%
Testing Data Results (N=341)						
Theme	**Domain Expert Coding**	**NLP Coding (NLP)**				
		LDA	***BERTopic***	***NMF***	***LSA***	***PCA***
Problem Solving	31.1%	16.4%	46.9%	27.0%	15.2%	30.5%
TA-Student Interactions	56.6%	58.9%	56.0%	64.5%	87.7%	64.8%
Experiential Learning	9.38%	24.6%	9.38%	21.4%	5.27%	16.1%

technique in both the training and testing data. These findings underscore the consistent performance of BERTopic, PCA, and NMF in accurately aligning with the domain expert's coding across various themes within the datasets.

From performance metric results, we found that the best match of NLP coding technique with domain expert coding was observed for the NMF and PCA techniques.

Given the strong performance of NMF and PCA, a more detailed evaluation was conducted using true positive, false positive, true negative, false negative, precision, recall, F1 score metrics and, Cohen's Kappa inter-rater reliability metric.

The accuracies with which each of the two NLP techniques (NMF, PCA) agreed with (i.e., assigned the same theme as) the human/domain expert are summarized in Table 4. Overall, PCA categorized student responses with accuracies ranging from 80.9% to 91.5% while NMF categorized those same responses with accuracies ranging from 80.0% to 86.3%.

In the Problem-Solving theme, NMF demonstrated a higher recall (78.0%) in the training data, indicating its ability to capture a larger proportion of positive instances. However, in the testing data, PCA showed a higher F1 score (70.4%), suggesting a better balance between precision and recall. For the TA-Student Interactions theme, NMF exhibited higher precision (89.6%) in the training data, while PCA demonstrated higher accuracy (82.4%) in the testing data. In the Experiential Learning theme, both PCA showed higher levels of accuracy in the training (89.1%) and testing datasets (91.5%) when compared to NMF (83.8% in training data and 86.2% in test data).

Table 4. Performance Metrics for NMF and PCA on Training and Testing Data (in percentage) Train Data (N=1359) and Test Data (N=341)

NMF									
Theme	**Data Type**	**TP**	**TN**	**FP**	**FN**	**Precision**	**Recall**	**F1 score**	**Accuracy**
Problem Solving	Train	23.5	60.0	9.79	6.62	70.6	78.0	74.1	83.5
	Test	19.0	61.0	7.92	12.0	70.6	61.3	65.6	86.3
Interactions with TAs	Train	43.7	42.6	5.08	8.54	89.6	83.6	86.5	85.9
	Test	48.3	35.4	9.38	6.74	83.7	87.7	85.7	80.0
Experiential Learning	Train	8.54	77.4	12.4	1.62	40.7	84.0	54.8	83.8
	Test	8.50	77.7	12.9	0.88	39.7	90.6	55.2	86.2
PCA									
Theme	**Data Type**	**TP**	**TN**	**FP**	**FN**	**Precision**	**Recall**	**F1 score**	**Accuracy**
Problem Solving	Train	20.1	60.7	9.05	10.0	69.0	66.8	67.9	80.9
	Test	21.7	60.1	8.80	9.38	71.1	69.8	70.4	81.8
Interactions with TAs	Train	48.4	32.4	15.2	3.83	76.1	92.6	83.5	80.9
	Test	49.2	33.1	11.7	5.87	80.7	89.3	84.8	82.4
Experiential Learning	Train	8.31	80.8	8.98	1.84	48.0	81.8	60.5	89.1
	Test	8.50	82.9	7.62	0.88	52.7	90.6	66.6	91.5

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)

A subsequent analysis of interrater reliability between the domain expert and NMF and PCA supported agreement between domain expert and NLP techniques (Table 5). Cohen's Kappa for all analyses ranged between 0.4 and 0.75 indicating that all interrater agreement was considered moderate (i.e., fair to good). Within this range though, there were some differences. For example, analyzing the TA support data with NMF resulted in a relatively high value of Cohen's Kappa (0.728) for TA-Student Interactions, indicating that NMF was particularly adept (i.e., bordering on excellent) at duplicating the domain expert's assignment for this particular theme.

In contrast, NMF was only fair (Cohen's Kappa = 0.477) at duplicating the Experiential Learning theme assigned by the domain expert.

Table 5. Cohen's Kappa (κ) for 2X2 analysis of Individual Themes within the Data

Theme	Training Data (N=1359)		Testing Data (N=341)	
	NMF	***PCA***	***NMF***	***PCA***
Problem Solving	0.622	0.544	0.517	0.573
TA-Student Interactions	0.728	0.614	0.672	0.640
Experiential Learning	0.477	0.548	0.485	0.622

The observed discrepancy between Cohen's Kappa and accuracy metrics might initially appear contradictory. However, this divergence can be attributed to the distinct characteristics of these measures. Cohen's Kappa accounts for the agreement that occurs by chance, offering a more nuanced understanding of model performance in relation to random guessing. In contrast, accuracy measures the direct correspondence between the model's classifications and the domain expert's coding without considering the probability of chance agreement. This distinction highlights the importance of considering both measures to gain a comprehensive understanding of the NLP techniques' performance.

Discussion

In this study, advanced Natural Language Processing (NLP) techniques were compared for analyzing student feedback regarding TA support in engineering education, aiming to uncover underlying patterns and insights that can inform and enhance pedagogical practices. The findings from our analysis provide valuable insight into how and when to use NLP in the analysis of student feedback in survey-based research.

Research Question #1 (RQ1):

How accurate are different NLP techniques in interpreting student feedback?

Previous research has highlighted the critical role of faculty and instructor interactions in student satisfaction with college education [35]. Consistent with these findings, our study identifies interactions with TAs as the most influential theme in student preferences within the context of engineering education. This suggests that TAs should prioritize engaging with students, although the importance of problem-solving support and experiential learning should not be overlooked.

A more detailed look at NLP-based performance metrics showed that, NMF and PCA achieved the highest average accuracy, nearly 84.0%, demonstrating their robustness and consistency [24], [27]. NMF's effectiveness may be attributed to its ability to uncover latent structures and patterns in data, facilitated by its non-negativity constraints that emphasize relevant information and its parts-based representations that simplify and enhance interpretability, effectively reducing noise and aligning closely with traditional coding methods [36]. PCA, known for its robustness to data variations and its capacity to manage noise, is particularly well-suited for analyzing textual data. By transforming the data into a set of orthogonal components that capture the most variance, PCA can distill the essence of the textual responses, enabling a more focused and interpretable analysis [27].

The performance metrics for NMF and PCA on both training and testing data offer valuable insights into the effectiveness of these topic modeling techniques compared to domain expert

coding [24], [27]. While NMF excels in capturing a larger proportion of positive instances, PCA provides a more balanced approach between precision and recall, particularly when applied to new, unseen data. This suggests that both NLP techniques have their unique strengths and can be effectively utilized for analyzing educational survey data, with PCA showing particular promise for handling large volumes of textual responses.

Research Question #2 (RQ2):
Can NLP replace domain expert coding in processing student feedback?
Answering this research question required a comparative analysis of the performance metrics employed in the NLP community. Our overarching conclusion is that NLP is best used in combination with a domain expert (i.e., hybrid approaches) in analyzing more complex data which is broad in scope and often ambiguous. Such is the case on short answer survey questions because the absence of assessment (grading) means that participants are more likely to venture afield in their responses and less likely to edit their responses are clear and unambiguous.

To assess the agreement between the domain expert and the NLP techniques, we used Cohen's Kappa, a measure that accounts for chance agreement. The analysis revealed moderate agreement (Kappa between 0.4 and 0.75) between the domain expert and both NMF and PCA for all themes explored (Table 5). Notably, NMF demonstrated a stronger ability to replicate the domain expert's coding for the TA-Student Interactions theme (Kappa = 0.728) compared to Experiential Learning (Kappa = 0.477). This suggests that NMF may be particularly effective at capturing themes related to student-instructor interactions, while both techniques require further refinement for specific themes like experiential learning.

Summary:
The wide range of interrater reliability agreement among the three themes and two NLP techniques used in this study prompted an in-depth analysis of why such discrepancies emerged in the classification success of each model. While analyzing ranking and Cohen's Kappa provides one lens (through that of traditional statistics and education research) to look at the performance of NLP models for classifying short answer survey data, other metrics specific to machine learning (including NLP) also provide some helpful insight into the goodness of each model.

However, observed discrepancies between Cohen's Kappa and accuracy metrics underscore the importance of using a multifaceted approach to evaluating NLP techniques. While accuracy provides a straightforward measure of the direct correspondence between NLP and domain expert coding, Cohen's Kappa offers a more nuanced assessment by accounting for the agreement that could occur by chance [37]. This distinction emphasizes the need for a comprehensive evaluation framework that considers both the precision of NLP techniques in replicating domain expert coding and their consistency in doing so across different themes and datasets.

Thus, the findings from this analysis suggest that while NMF and PCA exhibit considerable promise in automating the coding of educational survey data, particularly in areas where their strengths align closely with the thematic content of the responses, their application as replacements for domain expert coding should be approached with caution. Future research should continue to refine these NLP techniques, enhancing their accuracy and consistency, to

better harness their potential in educational settings. In particular, hybrid approaches that reduce but do not eliminate the role and time invested by a domain expert, are particularly promising for analyzing text-based data emerging from surveys and other educational instruments that fall outside of the assessment and evaluation of student performance.

Limitations

This study was conducted at a single, large research institution and the limited student population may have introduced biases caused by gender and racial demographics that were not necessarily representative of engineering student populations at other universities [21]. The under-representation and overrepresentation of certain groups of students in this study may impact the generalization of the results regarding TA support and is acknowledged as a limitation of this single institution study.

Additionally, the scope was confined to two engineering disciplines (mechanical and electrical engineering). Therefore, the themes and their relative importance may not translate directly to other engineering fields, especially those with a higher percentage of female students. While the themes of TA support identified in this study may resonate with other engineering student populations, their prioritization or significance could vary. However, the fact that the overall conclusions from both NLP and domain expert thematic analyses were the same indicates that even though the frequency with which these themes appear in the data might be different in another engineering student population, the accuracy of NLP is sufficient to be used in part as a substitute for resource-intensive traditional thematic coding conducted by a domain expert.

Implications

The implications of this study are twofold. Firstly, the findings underscore the potential of Natural Language Processing (NLP) techniques in capturing and categorizing student preferences and experiences related to Teaching Assistant (TA) support in engineering education. This suggests that NLP can be a valuable tool for educational researchers and practitioners, offering a more efficient and consistent method for analyzing large volumes of textual survey data. Secondly, the study highlights the potential for NLP techniques, particularly Non-Negative Matrix Factorization (NMF) and Principal Component Analysis (PCA), to complement or even replace domain expert coding for categorizing short-answer responses. The high levels of agreement between NMF and PCA with domain expert coding, as indicated by Cohen’s Kappa analysis, suggest that these NLP techniques can offer reliable and consistent results, potentially reducing the need for extensive manual coding in educational research. These implications are important for informing future research and practice in the field of educational data analysis.

The study's findings also have implications for policy, practice, theory, and subsequent research. For instance, the consistent trends observed across different NLP techniques and manual coding by domain experts indicate the reliability of hybrid NLP methods in identifying key themes in student responses. This suggests that NLP can be a valuable tool for educational researchers and practitioners, offering a more efficient and consistent method for analyzing large volumes of textual survey data. Additionally, the high levels of agreement between NMF and PCA with domain expert coding suggest that these NLP techniques can offer reliable and consistent results, potentially reducing the need for extensive manual coding in educational research. These

implications are important for informing future research and practice in the field of educational data analysis.

Conclusions

In this study, we explored the application of Natural Language Processing (NLP) techniques for automating the coding and analysis of student feedback regarding Teaching Assistant (TA) support in engineering courses. Our findings demonstrate the efficacy of hybrid NLP techniques (that use both automated and manual approaches to analyzing data) in capturing student preferences and experiences, offering insights into TA pedagogy and practice. Non-Negative Matrix Factorization (NMF) and Principal Component Analysis (PCA) are particularly promising for replacing a large portion of domain expert coding and thematic analysis. Both NMF and PCA demonstrated high levels of agreement with domain expert coding, as indicated by Cohen's Kappa analysis. Additionally, NMF exhibited higher recall rates in capturing positive instances, while PCA showed better precision and overall balance between precision and recall.

Moving forward, further research is necessary to refine these NLP techniques for educational contexts and to optimize the role of the domain expert in the hybrid approach. Additionally, ethical considerations surrounding the use of NLP in educational research, such as student privacy and potential biases within algorithms, should be addressed in future work. This paper, however, has laid additional groundwork for implementing NLP techniques in educational research on a broad scale.

Acknowledgements

The authors would like to gratefully acknowledge the National Science Foundation for their partial support of this work (DUE grant number 1504618). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] National Center for Education Statistics. (2020). *The SAGE Encyclopedia of Higher Education*. https://doi.org/10.4135/9781529714395.n400

[2] M. Parry (2012). " Supersizing" the College Classroom: How One Instructor Teaches 2,670 Students. *Chronicle of Higher Education*.

[3] M. Soledad, J. Grohs, S. Bhaduri, J. Doggett, J. Williams, and S. Culver, "Leveraging institutional data to understand student perceptions of teaching in large engineering classes," *2017 IEEE Frontiers in Education Conference (FIE)*, Oct. 2017. https://doi.org/10.1109/fie.2017.8190608

[4] E. Blair and K. Valdez Noel, "Improving higher education practice through student evaluation systems: is the student voice being heard?," Assessment & *Evaluation in Higher Education*, vol. 39, no. 7, pp. 879–894, Jan. 2014, doi: 10.1080/02602938.2013.875984.

[5] G. G. Chowdhury, "Natural language processing," *Annual Review of Information Science and Technology*, vol. 37, no. 1, pp. 51–89, Jan. 2005, doi: 10.1002/aris.1440370103.

[6] R. S. Baker and P. S. Inventado, "Educational Data Mining and Learning Analytics," *Learning Analytics*, pp. 61–75, 2014, doi: 10.1007/978-1-4614-3305-7_4.

[7] N. Kardam, S. Misra, and D. Wilson, "Is Natural Language Processing Effective in Education Research? A case study in student perceptions of TA support," presented at the *2023 ASEE Annual Conference & Exposition, 2023*. [Online]. Available: https://peer.asee.org/43887

[8] 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

[9] D. G. Oblinger, "Let's Talk... Analytics," *Educause Review*, vol. 47, no. 4, pp. 10-13, 2012.

[10] J. P. Magliano and A. C. Graesser, "Computer-based assessment of student-constructed responses," *Behavior Research Methods*, vol. 44, no. 3, pp. 608–621, May 2012, doi: 10.3758/s13428-012-0211-3.

[11] 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.

[12] D. Wang, J. Su, and H. Yu, "Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language," *IEEE Access*, vol. 8, pp. 46335–46345, 2020, doi: 10.1109/access.2020.2974101.

[13] S. Gombert et al., "From the Automated Assessment of Student Essay Content to Highly Informative Feedback: A Case Study," *International Journal of Artificial Intelligence in Education*, Jan. 2024, doi: 10.1007/s40593-023-00387-6.

[14] A. Bralin, J. W. Morphew, C. M. Rebello, and N. S. Rebello, "Analysis of student essays in an introductory physics course using natural language processing," *2023 Physics Education Research Conference Proceedings*, Oct. 2023, doi: 10.1119/perc.2023.pr.bralin.

[15] Kerkhof, R. G. (2020, June). Natural Language Processing for Scoring Open-Ended Questions: A Systematic Review. [Online]. Available: http://essay.utwente.nl/82090/

[16] V. S. Sadanand, K. R. R. Guruvyas, P. P. Patil, J. Janardhan Acharya, and S. Gunakimath Suryakanth, "An automated essay evaluation system using natural language processing and sentiment analysi," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, p. 6585, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6585-6593.

[17] F. Dalipi, K. Zdravkova, and F. Ahlgren, "Sentiment Analysis of Students' Feedback in MOOCs: A Systematic Literature Review," *Frontiers in Artificial Intelligence*, vol. 4, Sep. 2021, doi: 10.3389/frai.2021.728708.

[18] E. Mayfield, M. Madaio, S. Prabhumoye, D. Gerritsen, B. McLaughlin, E. Dixon-Román, and A. W. Black, "Equity beyond bias in language technologies for education," in *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, 2019, pp. 444-460. https://doi.org/10.18653/v1/w19-4446

[19] N. Arthurs and A. J. Alvero, "Whose Truth Is the 'Ground Truth'? College Admissions Essays and Bias in Word Vector Evaluation Methods," *International Educational Data Mining Society*, 2020.

[20] 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.

[21] "Engineering and Engineering Technology by the Numbers, 2021," *American Society for Engineering Education (ASEE), 2021*. [Online]. Available: https://ira.asee.org/wp-content/uploads/2022/09/Engineering-and-Engineering-Technology-by-the-Numbers-2021.pdf. [Accessed: Feb. 06, 2024].
[22] 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].
[23] D. M. Blei and M. I. Jordan, "Variational methods for the Dirichlet process," *Twenty-first international conference on Machine learning - ICML* '04, 2004, doi: 10.1145/1015330.1015439.
[24] N. Gillis, "The why and how of nonnegative matrix factorization," *Connections*, vol. 12, no. 2, 2014, doi: 10.1137/1.9781611976410.
[25] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.
[26] S. T. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 188–230, Sep. 2005, doi: 10.1002/aris.1440380105.
[27] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jun. 2010, doi: 10.1002/wics.101.
[28] S. F. Chen, D. Beeferman, and R. Rosenfeld, "Evaluation Metrics For Language Models," *Carnegie Mellon University*, 2018. [Online]. Available: https://doi.org/10.1184/R1/6605324.v1. [Accessed: Feb. 06, 2024].
[29] "Perplexity," *Wikipedia*, Jan. 28, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Perplexity. [Accessed: Feb. 06, 2024].
[30] "What is Perplexity in NLP," *Educative Answers*, Jan. 29, 2024. [Online]. Available: https://www.educative.io/answers/what-is-perplexity-in-nlp. [Accessed: Feb. 06, 2024].
[31] N. Gisev, J. S. Bell, and T. F. Chen, "Interrater agreement and interrater reliability: key concepts, approaches, and applications," *Research in Social and Administrative Pharmacy*, vol. 9, no. 3, pp. 330-338, Sep. 2013. doi: 10.1016/j.sapharm.2012.04.004.
[32] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 1, 2015.
[33] T. F. Monaghan, S. N. Rahman, C. W. Agudelo, A. J. Wein, J. M. Lazar, K. Everaert, and R. R. Dmochowski, "Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value," *Medicina (Kaunas)*, vol. 57, no. 5, p. 503, May 2021. DOI: 10.3390/medicina57050503.
[34] 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].
[35] A. W. Astin, "Student involvement: A developmental theory for higher education," *Journal of College Student Personnel*, vol. 25, no. 4, pp. 297-308, 1984
[36] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: A comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, p. 42, 2020. [Online]. Available: https://doi.org/10.3389/frai.2020.00042

[37] A. S. Kolesnyk and N. F. Khairova, "Justification for the Use of Cohen's Kappa Statistic in Experimental Studies of NLP and Text Mining," *Cybernetics and Systems Analysis*, vol. 58, pp. 280–288, 2022. [Online]. Available: https://doi.org/10.1007/s10559-022-00460-3