The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center ASEE
Portland, OR . June 23 - 26, 2024
Paper ID #44064

A Hybrid Approach to Natural Language Processing for Analyzing Student Feedback about Faculty Support

Neha Kardam, University of Washington

Neha Kardam is a fourth-year Ph.D. student 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.

A Hybrid Approach to Natural Language Processing for Analyzing Student Feedback about Faculty Support

Abstract

Short-answer questions in surveys serve as a valuable educational tool, used for evaluating student learning and exploring the perspectives of various stakeholders in educational research. However, it is essential to distinguish between the objectives of automated short answer scoring systems (ASAS) and automated short answer coding (ASAC) systems. ASAS aims to achieve high accuracy primarily for fair assessment, while ASAC systems can accommodate slightly lower accuracy without compromising the validity of conclusions drawn from code analysis in the context of survey responses.

This study focuses on a dataset comprising responses from 1857 undergraduate students who were asked to express their views on how faculty could enhance their learning experiences. The dataset encompasses students from different engineering majors and various learning environments, presenting a challenge due to its intricacy and variability. To address this challenge, the study introduces a novel approach by integrating domain expert interaction and unsupervised learning into the ASAC process, specifically tailored for complex and heterogeneous datasets. Unlike previous research, this study emphasizes the iterative process of code refinement by domain experts, which is a significant departure from fully automated methods.

By combining expert insights with non-negative matrix factorization (NMF), our research demonstrates that unsupervised learning can achieve accuracy comparable to supervised learning for complex qualitative data. This finding suggests a new paradigm where domain expertise can significantly enhance the performance of machine learning techniques in educational research.

The study highlights the importance of domain expert interaction in creating a robustly labeled dataset for the faculty support survey question. This interaction was critical in achieving high precision, recall, and F1 scores (91.3% to 99.3%) in supervised learning models, indicating a successful prediction of themes in student responses.

Our research provides evidence that the integration of expert knowledge can improve the efficiency and effectiveness of ASAC systems. It also demonstrates the potential to replace manual coding with automated NLP coding, supported by moderate to substantial agreement (Cohen's kappa) between expert raters and between expert and NLP coding.

This study not only presents a methodological innovation by merging expert interaction with unsupervised learning but also provides empirical evidence of its effectiveness, offering a valuable contribution to the field of educational research and the development of ASAC systems.

Introduction

In the dynamic landscape of education research, the evaluation of student experiences and perspectives is integral for fostering effective learning environments. Short answer questions in surveys and assessments provide valuable insights, capturing student perspectives on support, learning outcomes, and satisfaction. Traditional qualitative methods, while valuable for their depth and nuance, often struggle to efficiently handle the vast amount of textual data generated by student surveys and assessments [1]. This data, typically collected through short answer questions, offers rich insights into student perceptions of educational support, learning outcomes, and overall satisfaction. Automated Short Answer Coding (ASAC) systems have emerged as a promising solution for processing these large-scale qualitative data. However, the inherent complexity and heterogeneity of student responses pose significant challenges for achieving accurate automated analysis [1].

The limitations of traditional qualitative methods in educational research have been well-documented [2]. Studies highlight the challenges associated with manual coding, particularly the time-intensive nature of the process and the potential for subjectivity in interpretation [2]. While automated text analysis is a new and promising area within education research, a gap exists between qualitative and quantitative approaches [3]. This gap presents an opportunity to leverage the strengths of both methodologies to extract deeper meaning from student feedback data.

Machine learning offers a powerful set of tools to bridge this gap. Unsupervised learning techniques, such as Non-negative Matrix Factorization (NMF), hold significant promise for analyzing large volumes of unstructured text data in educational research [4]. The ability of unsupervised learning to identify latent themes without pre-defined categories makes it particularly well-suited for the inherent diversity of student perspectives [5]. Supervised learning techniques, such as Naïve Bayes and Support Vector Machines (SVMs), can then build upon the initial insights gleaned from unsupervised learning to refine and predict thematic structures within student responses [5].

This research delves into a novel approach that integrates domain expert interaction with both unsupervised and supervised learning techniques within the ASAC process. The study seeks to address two key methodological research questions: (1) How can the effective integration of domain expert interaction enhance the accuracy and efficiency of unsupervised learning for analyzing student feedback data? (2) To what extent does the inclusion of domain expert interaction in the unsupervised learning stage impact the subsequent application of supervised learning techniques for theme prediction? By exploring these questions, this study aims to contribute to the development of educational research methodologies. It emphasizes the potential collaboration between automated coding systems and human expertise in interpreting student feedback data.

Literature Review

Over 16 million people are enrolled as undergraduates in colleges and universities in the US [6]. Understanding the lived experiences of these students on a broad scale including their satisfaction with their education, learning outcomes, and intentions to persist in their careers requires education-based research that extends beyond the standard Likert-scale questions on surveys and student evaluations of teaching [1]. Augmenting surveys with short answer questions allows researchers and instructors to more effectively and more thoroughly interpret student feedback on course outcomes, instructional support, and other facets of both informal and formal learning. Unfortunately, analyzing short answer responses on surveys and other instruments is very time-consuming using traditional thematic and other qualitative analysis methods when conducted by (human) domain experts [7].

In contrast to automatic short answer scoring (ASAS) systems [8], which evaluate student learning, automatic short answer coding (ASAC) systems focus on assessing student perceptions of their educational experiences. ASAC systems encounter the challenge of dealing with a diverse range of student responses, which can be highly heterogeneous. However, unlike ASAS systems, ASAC systems are not obliged to achieve the exceptionally high accuracies necessary for fairness and equity in evaluating student understanding. Lower accuracies on a per-item basis are deemed acceptable for ASAC systems, as long as the overall conclusions drawn align with those derived from traditional coding and analysis methods. Thus, dramatic reductions in the time and effort required to process short answer, text-based data in education research are indeed possible using computer-assisted ASAC methods [2].

Despite the potential of ASAC methods, a wide range of studies, have emphasized that using standalone ASAC systems, when used as the sole means to analyze educational research data and evaluate the results, is insufficient. Rather, domain expert interaction in the ASAC process has proven useful to increasing the value of automated data analysis [9]. Domain expert interaction refers to the involvement of a human being in ASAC who is both experienced in qualitative data analysis and in the educational research domain associated with a particular dataset. Domain expert interaction with ASAC can occur anywhere in the data analysis process – in data cleaning, data preprocessing, topic modeling and formulation, topic aggregation and theme building, etc.

One approach is to engage domain experts in ASAC is to include them in the optimization of the number of topics represented by the data. For example, in topic modelling, domain experts might choose to combine topics that emphasize “Russia” and “Soviet Union” into a single topic, or they may choose to split a topic which they deem to contain mixed content. Hu et al. (2013) leveraged this approach directly into the modelling process itself, using correlations to modify naïve topic modelling algorithms into more informed, “intelligent” tree-based language models [5].

Other researchers have used ASAC to support traditional qualitative (thematic) data analysis. Katz et al. (2021) used NLP-based ASAC to generate optimal numbers of topics (codes) to represent the short-answer experiences of engineering students during the COVID-19 pandemic [9]. These codes were then aggregated by a domain expert to create themes for subsequent thematic analysis and to enable a more nuanced evaluation of themes across the dataset. Similarly, Fernandez et al. (2021) also involved domain expert interaction at the back end of analyzing the short-answer responses of teachers to prompts about their retention strategies [10]. The researchers used ASAC methods including topic modeling and text network modeling algorithms to identify clusters of related topics and visualize the connections between them [10]. The domain expert then reviewed and further refined the topics identified by ASAC to create more meaningful and interpretable themes. Domain expert interaction has also been used midstream in qualitative data analysis. Zhang et al. (2019) used ASAC involving word2vec to create word embeddings that capture semantic similarity between words within short answer data [11]. They then clustered the data based on these word embeddings to identify the most common topics expressed by respondents. Midstream in the analysis, domain experts then manually labeled a subset of the survey responses for each topic. At the end of the analysis, these labeled responses were then used to train a logistic regression model which automatically predicted topic labels for new survey responses [11].

Regardless of where in the data analysis process domain expert interaction occurs, it has been shown to be a valuable and integral part of fulfilling the potential of ASAC to improve the efficiency of educational research and to expand the use of qualitative methods in such research [9]. In this paper, we further contribute to the existing knowledge base regarding domain expert interaction by investigating the value of (a) unsupervised (Non-negative Matrix Factorization) learning techniques that use various degrees of domain expert interaction and (b) supervised (naïve Bayes and support vector machine) for analyzing short answer responses from a diverse survey dataset. The dataset is a highly heterogeneous (and therefore challenging) collection of text-based data, generated from asking a "reach for the moon" question to students about what they would like faculty to do to better support their learning.

Methods

The study was conducted at a large public research institution located in an urban setting to explore various forms of instructional support and course level engagement such as faculty support, student-faculty interactions, attention, participation, effort, and emotional engagement. The survey contained both open- and close-ended questions One of the open-ended questions focused on faculty support and asked students to respond to the question: "*What one action can your faculty at <this institution> take to best support you in your classes (please be as specific as possible)?".* To analyze responses to the qualitative survey on faculty support this question, NLP was used, incorporating domain expert input and interaction to automate, in whole or in part, coding and qualitative analysis. Two research questions were formulated to guide this process:

Methodology Research Question (RQ1):
How can domain expert interaction be effectively integrated into the automatic short answer coding (ASAC) and thematic analysis process to enhance the value of NLP in educational research?

Methodology Research Question (RQ2):
How does data analysis involving both unsupervised learning methods compared with analysis using only unsupervised methods?

Participants

The study involved a total of 1,857 participants, consisting of sophomores and juniors from four different engineering majors enrolled as undergraduates. The participants were surveyed between the winter of 2017 and the spring of 2022. The study population was divided into two settings: traditional (in-person) prior to the COVID-19 pandemic and emergency remote teaching (ERT), which was conducted remotely during the pandemic. The gender distribution showed that 74.1% of the participants were male, 24.4% were female, and a small percentage identified as "Other." In terms of race, the majority of participants were White (37.7%), followed by Asian American (19.2%) and Asian International (11.7%). The U.S. status of the participants indicated that 82.2% were domestic students, while 17.7% were international. Detailed demographics across both traditional and ERT time periods are summarized in Table 1.

Table 1. Demographics of study population (*N* = 1,857)

Demographic Variable	*N*	%	*N*	%	*N*	%
	All Students		Traditional Setting		ERT Setting	
Total	1857	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

Procedure

The study was approved by the Internal Review Board (IRB) under the code STUDY00000378. The study recruited undergraduate students from 21 courses in mechanical and electrical engineering, but the researchers did not engage directly with the students. All participants were informed that their responses would be kept confidential. Additional academic incentives, in the form of extra credit, were provided to students to support increased survey participation and all surveys were conducted electronically.

Data Analysis

Raw data from student responses was initially processed using Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer to convert the unstructured data into structured format [12]. The TF-IDF Vectorizer provided by Sklearn.org calculates a score that reflects the relevance of a term in a document in relation to its significance across the entire corpus [13]. Terms (words) that appear more frequently in a particular document (i.e., student response) but rarely across the entire corpus (i.e., dataset) have higher scores, while terms in a document that appear more frequently in the response of other students receive a lower TF-IDF score [14], [15].

During vectorization, we employed specific hyperparameters to refine the process:

- $max\_df = 0.8$: This parameter ignores terms with a document frequency exceeding the threshold (here, 0.8). This helps remove very frequent words that may not be informative for the analysis. For example, with a document frequency of 79.6%, the word "class" would likely be excluded using this setting.
- $min\_df = 2$: This parameter ignores terms appearing in less than the specified number of documents (here, 2). This removes rare words that may not contribute significantly to the analysis. For instance, the word "Member" (document frequency of 26.8%) might be excluded with this setting.

- $ngram\_range = (1,1)$: This parameter specifies that only unigrams (single words) are considered for analysis. Changing this to (1,2) would include bigrams (two-word phrases) as well.
- $stop\_words = 'english'$: This parameter removes common English stop words (e.g., "the," "is," "in") from the text data, as these words generally don't provide significant meaning for modeling [12].

Applying these hyperparameters allowed us to focus on terms that are informative within individual responses while also exhibiting variation in usage across the dataset. Notably, max_df and min_df don't directly remove words from the vocabulary. Instead, the vectorizer creates a vocabulary of all encountered terms. However, these parameters influence the TF-IDF scores assigned to each term. Terms with very high or very low document frequencies receive low weights (TF-IDF scores), minimizing their impact on the subsequent Non-negative Matrix Factorization (NMF) model. This focus on informative and varied terms ultimately contributes to a more meaningful analysis using an unsupervised learning method known as Non-negative Matrix Factorization (NMF), domain expert interaction and supervised learning methods. Three different approaches (methods) to analyzing the data are explained below in Figure 1.

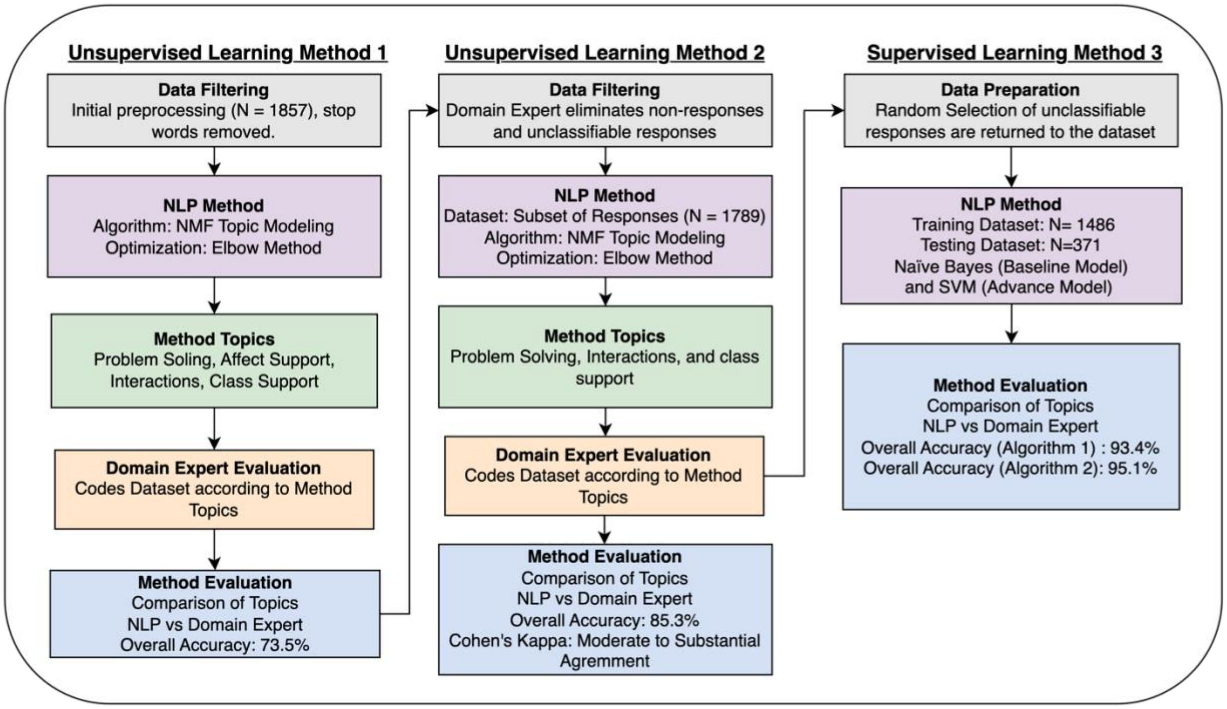

Unsupervised Learning Method 1
Data Filtering
Initial preprocessing (N = 1857), stop words removed.
NLP Method
Algorithm: NMF Topic Modeling
Optimization: Elbow Method
Method Topics
Problem Soling, Affect Support, Interactions, Class Support
Domain Expert Evaluation
Codes Dataset according to Method Topics
Method Evaluation
Comparison of Topics
NLP vs Domain Expert
Overall Accuracy: 73.5%
Unsupervised Learning Method 2
Data Filtering
Domain Expert eliminates non-responses and unclassifiable responses
NLP Method
Dataset: Subset of Responses (N = 1789)
Algorithm: NMF Topic Modeling
Optimization: Elbow Method
Method Topics
Problem Solving, Interactions, and class support
Domain Expert Evaluation
Codes Dataset according to Method Topics
Method Evaluation
Comparison of Topics
NLP vs Domain Expert
Overall Accuracy: 85.3%
Cohen's Kappa: Moderate to Substantial Agremment
Supervised Learning Method 3
Data Preparation
Random Selection of unclassifiable responses are returned to the dataset
NLP Method
Training Dataset: N= 1486
Testing Dataset: N=371
Naïve Bayes (Baseline Model) and SVM (Advance Model)
Method Evaluation
Comparison of Topics
NLP vs Domain Expert
Overall Accuracy (Algorithm 1) : 93.4%
Overall Accuracy (Algorithm 2): 95.1%

Figure 1: Research Method Flow Diagram

Unsupervised Learning, Method #1: The pre-processed data was analyzed using NMF without any additional preprocessing or filtering as illustrated in Figure 1. NMF is a widely used and effective method for extracting topics from textual data to uncover latent themes [16]. NMF operates on a document-term matrix whose rows correspond to the number of documents (e.g., student responses) in a dataset and columns correspond to the raw count of terms or a pre-processed version of that raw count (e.g., TF-IDF score). NMF decomposes the document-term matrix of textual data into two non-negative matrices, one matrix reflects the topics within documents (the W matrix) and the second the distribution of terms within topics (the H matrix) [16]. The initial selection of W and H matrices is random and NMF iterates until a cost function (e.g., Frobenius norm) is minimized. The resulting W and H matrices uncover the key topics and terms associated with those topics in the dataset.

Using NMF, Method #1 determines the optimal number of topics that best describes the data by selecting the lowest value of perplexity across 2, 4, 5, 6, and 7 topics. The lowest perplexity score ensures that the model provides a more accurate representation of the underlying topics within the given dataset [17]. Once the optimal number of topics is determined, the terms (words) for each topic are visualized using word clouds and a domain expert uses these word clouds to describe the themes embodied by the topics. A single topic may comprise a single theme or multiple topics may be combined into a single theme by the domain expert. Regardless, once the themes are determined and descriptions for teach theme manually generated, the domain expert uses the themes and corresponding descriptions to manually designate a theme for each student response. These themes are then used as a ground truth for comparison with NLP assigned codes. Agreement or lack of agreement between the ground truth and the themes generated by NLP is then used as a measure of the success of Method #1.

Unsupervised Learning, Method #2: Method #1 presumes that the converting the raw data to lower case, eliminating stop words, and TD-IDF vectorization are sufficient for pre-processing the data. Method #2 added additional pre-processing by relying on the domain expert to remove responses that did not address the question/prompt regarding faculty support either because the student stated that they had no suggestions to offer, answered in a way that was not related to the question, or articulated in such a way that their response was “ambiguous" and not be categorized into any theme. The pre-processed data was then processed as in Method #1 using NMF to identify and determine the optimal number of topics in the dataset.

Supervised Learning, Method #3: Responses coded as "No response, other, or ambiguous" by the domain expert in Method #1 were added in random order to the data from Method #2. This pre-processed data was used as input for Supervised Learning Method #3. The preprocessed data was randomly split into two sets at an 80:20 ratio, trained on the larger data subset and tested on the smaller subset. Method #3 utilized two supervised learning algorithms, namely Naïve Bayes and Support Vector Machine (SVM). The Naïve Bayes algorithm is a popular and widely used probabilistic classifier which assumes independence between features, making it easy to implement and interpret [18]. Naïve Bayes served as the baseline model for supervised learning, allowing us to establish a performance benchmark against which we could compare support vector machine approaches to analyze the data. Support Vector Machine (SVM) served as the more advanced supervised learning model. SVM is a powerful and versatile algorithm that separates data into classes by finding the best hyperplane that maximally separates the data points. It is particularly useful when dealing with high-dimensional data and has been shown to outperform other algorithms in many classification tasks [19].

Methods #1 and #2 were used to evaluate the first research question associated with this study (RQ1) and Methods #2 and #3 were used to evaluate the second research question (RQ2). Domain expert (manual) coding results are compared to NLP (automated) coding results using a number of performance metrics including True Positives (TP), True Negatives (TN), False

Positives (FP), False Negatives (FN), and overall accuracy [20]. A true positive (TP) occurs when the domain expert and the NLP method associate a student's response with the same theme while a false positive (FP) occurs when the NLP does not assign a theme to a response that the domain expert does. True Negatives (TN) represent the number of responses correctly identified as not pertaining to a specific theme by both domain expert and NLP method and False Negatives (FN) refer to instances where the automated system fails to associate a theme with a student's response when the domain expert does do so. In addition to these performance metrics, Cohen's kappa is used to measure agreement between NLP and domain expert assigned themes for Methods #1, #2, and #3. Cohen's kappa is a statistical measure that assesses the level of agreement between domain expert and NLP coding [21], [24]. Intercoder reliability was also assessed to identify potential biases in manual (domain expert) coding by identifying the level of agreement between the primary domain expert and a secondary domain expert who coded a random subset of responses independently from the primary domain expert.

To further address RQ2, precision, recall, and F1-score of the supervised learning methods were also assessed. Precision measures the proportion of true positive predictions out of all positive predictions, while recall measures the proportion of true positive predictions out of all actual positives [20], [23]. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the predictive model's accuracy [20], [23].

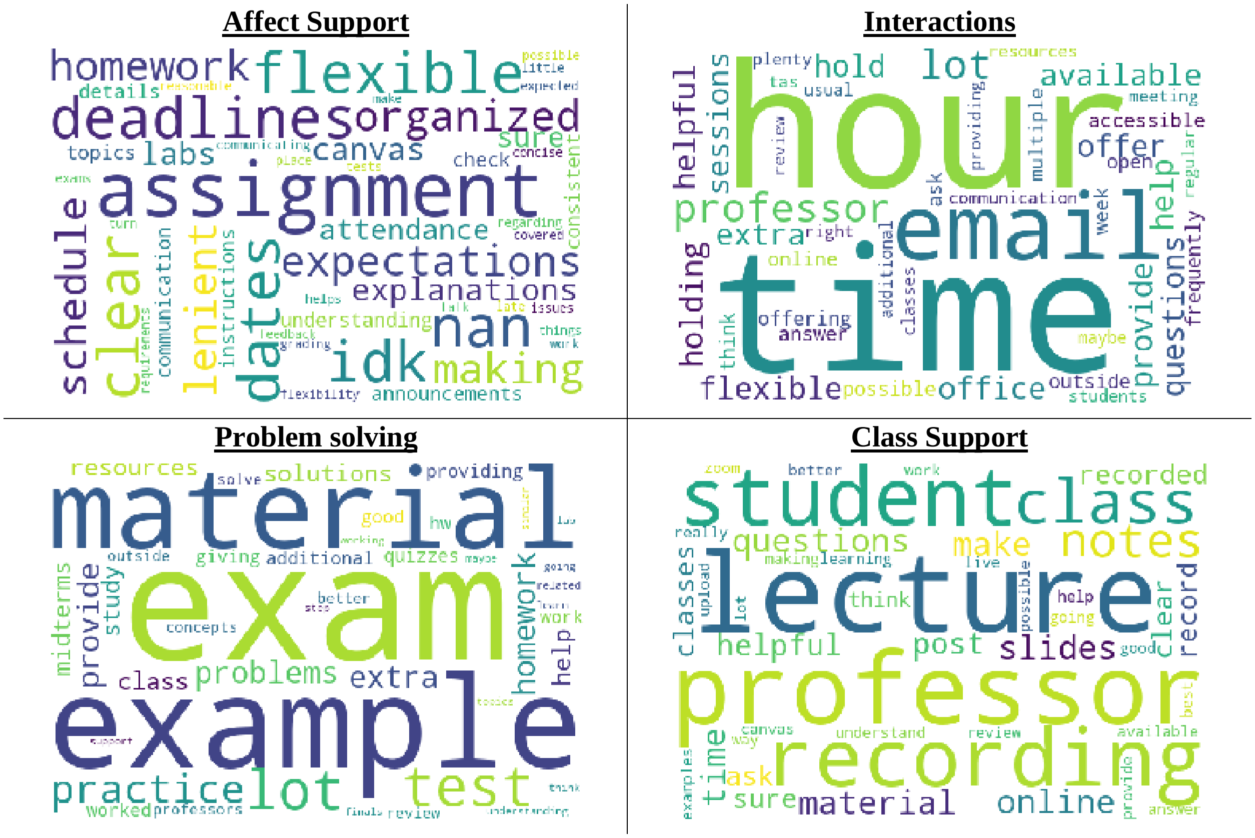
Results

In this study, student responses to a short answer question regarding their preferences for faculty support were subjected to preprocessing as outlined in the data analysis step. The preprocessed data was then analyzed using unsupervised learning method 1, followed by further preprocessing in unsupervised learning method 2, and finally in supervised learning method 3. Below are the results of each method used in this study.

Unsupervised Learning, Method 1: To serve as a baseline for understanding how domain expert interaction can be used to assist ASAC in education research, NMF-based topic modelling was applied to the entire data set in a fully automated approach to coding the data. Four topics, corresponding to the lowest perplexity score in NMF, were chosen to be optimal and the resulting word clouds for these four topics are shown in Figure 2. The domain expert evaluated these word clouds, created labels (codes), and derived code descriptions to guide subsequent coding of the data. Shortened summaries of these codes are described below.

- Topic 1 (coded *Affect Support*): relating to student needs for understanding, flexibility, lenience, and other affective support of their learning.
- Topic 2 (coded *Class Support*): relating to activities and resources necessary to prepare for class and deliver lectures.
- Topic 3 (coded *Problem Solving*): relating to resources, pedagogy, and other instructional activities that focus on practicing the application of course concepts to engineering problems.
- Topic 4 (coded *Interactions*): relating to interactions between faculty and students, TAs and students, as well as among students.

To evaluate the practical value of Method 1, the domain expert assigned a code to each response in the dataset using these code descriptions. The results in Table 2 provide a comprehensive insight into the performance of the NLP-assigned themes in comparison to those assigned by a domain expert.


Affect Support

Interactions

Problem solving

Class Support

Figure 2: Word Clouds representing the four topics emerging from Method #1 ASAC

For the "Affect Support" theme, the model demonstrates a notable accuracy of 89.2%, primarily driven by a high level of true negatives (89.2%). However, the challenge lies in the low proportion of true positives (0.09%), indicating the model's struggle to accurately identify instances of "Affect Support." In comparison, the "Class Support" theme exhibits a more balanced distribution, with 37.8% true positives and 27.3% true negatives.

Both the "Interactions with TA’s" and "Problem Solving" themes exhibit similar patterns of high true negatives (74.4% for "Interactions with TA’s" and 73.1% for "Problem Solving") and accuracy (85.1% for "Interactions with TA’s" and 54.7% for "Problem Solving"). The domain expert also discovered that 70 responses did not fit into any of the four categories and were thus classified as "No response, other, or ambiguous".

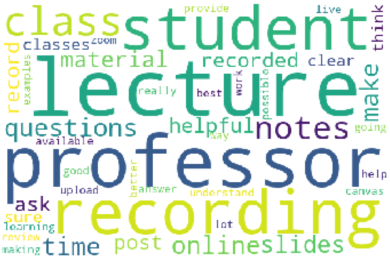
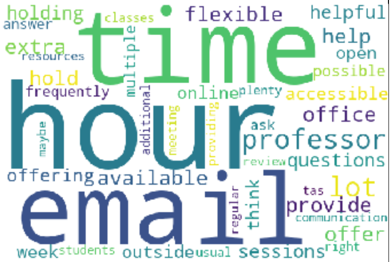
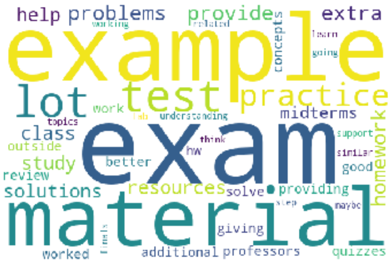
Table 2. Comparison of NLP Assigned Themes to those Assigned by a Domain Expert (N=1855)

Theme	True Positives	True Negatives	False Positives	False Negatives	Accuracy
Affect Support	0.09%	89.2%	2.80%	8.00%	89.2%
Class support	37.8%	27.3%	26.6%	8.30%	65.1%
Interactions	10.7%	74.4%	2.00%	13.0%	85.1%
Problem Solving	10.1%	73.1%	9.90%	6.80%	54.7%

Unsupervised Learning, Method 2:

The domain expert's review of the data in the previous method revealed the presence of non-answers, other, and ambiguous responses in many instances. Consequently, these responses were removed from the dataset based on the domain expert's interaction with the data, resulting in 1785 responses remaining for analysis. Subsequent ASAC using NMF-based topic modeling of the reduced dataset led to a low perplexity score for three topics. Table 3 displays the resulting topic, theme, top ten most frequently associated words, and word clouds for these three topics. These findings shed light on the model's performance in capturing the specified themes within the dataset, emphasizing the need for a nuanced evaluation of its effectiveness across different thematic categories.

Table 3. Topics, Themes, and word clouds emerged from Method #2
(Most Frequently Occurring Words associated with Each Topic (N=1785))

Topic 1	**Topic 2**	**Topic 3**
'lecture', 'class', 'lectures', 'students', 'notes', 'questions', 'time', 'make', 'online', 'slides'	'office', 'hours', 'hold', 'available', 'offer', 'extra', 'help', 'provide', 'having', 'hour'	'practice', 'problems', 'exams', 'provide', 'examples', 'tests', 'homework', 'extra', 'exam', 'example'
Theme 1 Class support 	**Theme 2** Interactions 	**Theme 3** Problem Solving 

To evaluate the practical value of Method 2, the same technique was used as for Method 1. The domain expert coded the topics, derived code descriptions, and then coded the data accordingly. The resulting codes are abbreviated code descriptions are:

- Topic 1 (coded *Class Support*): relating to activities and resources necessary to prepare for class and deliver lectures.
- Topic 2 (coded *Problem Solving*): relating to assessment (e.g., homeworks, exams), resources, pedagogy, and other instructional activities that focus on practicing the application of course concepts to engineering problems.
- Topic 3 (coded *Interactions)* relating to interactions between faculty and students, TAs and students, as well as among students outside of class.

Table 4. NLP Assigned Themes Vs Domain Expert Assigned Themes for Method #2

Theme	**True Positives**	**True Negatives**	**False Positives**	**False Negatives**	**Accuracy**
Class support	48.5%	31.3%	15.7%	4.49%	80.0%
Interactions	12.9%	73.7%	1.35%	12.1%	86.5%
Problem Solving	16.5%	72.9%	5.05%	5.05%	89.4%

Table 4 presents performance metrics for unsupervised learning Method 2. Both NLP Coding (NMF) and Domain Expert Coding demonstrated the highest accuracy in identifying "Problem

Solving (89.4%)," followed by "Interactions (86.5%)" and "Class Support (80.0%)."Class Support" has the highest False Negative rate, suggesting that the NLP model struggled more in identifying instances of "Class Support" compared to the other themes. However, it is performing well in identifying true positives (48.5%) when compared to other themes. To gauge into topic/themes, we looked at the agreement between NLP coding and domain expert coding in interpreting the three themes.

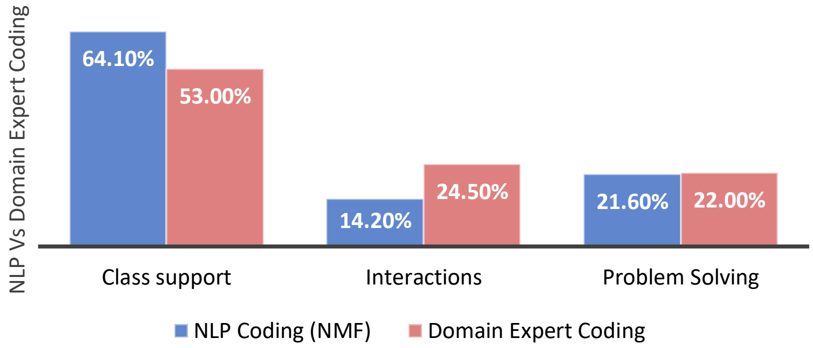


Figure 3: Percentage Agreement: NLP Domain VS Expert Coding

In Figure 3, it is evident that both coding methods identified the topics of "Class Support" and "Problem Solving" as the most frequent forms of instructional support desired by students. The NLP Coding (NMF) and Domain Expert Coding showed a high level of agreement in categorizing student responses into these themes, with "Class Support" and "Problem Solving" However, there was a notable difference in the identification of the "Interactions" theme, with NLP Coding (NMF) assigning this theme to 14.2% of student responses, while the Domain Expert Coding identified it in 24.5% of responses. This difference suggests a potential discrepancy in the interpretation of student responses related to interactions with teaching assistants between the two coding methods.

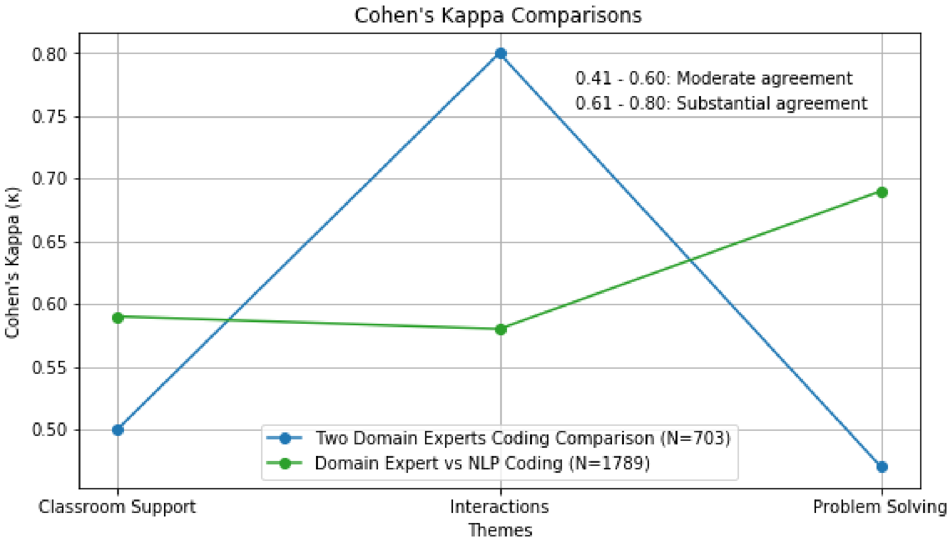


Figure 4: Cohen's Kappa (κ) for inter-rater reliability

The Cohen's Kappa coefficient, which is used to assess inter-rater reliability or agreement, provided further insights into the level of agreement between the NLP Coding (NMF) and Domain

Expert Coding as well as an inter-rater reliability test between two domain expert coders [21]. The results summarized in Figure 4, indicate varying levels of agreement between raters. For "Classroom Support," there is moderate agreement between two domain experts (κ = 0.50) and a slightly higher, yet still moderate, agreement between domain expert and NLP coding (κ = 0.59). "Interactions" show substantial agreement among domain experts (κ = 0.80), but this drops to moderate when compared with NLP coding (κ = 0.58). "Problem Solving" has a moderate agreement between domain experts (κ = 0.47) and substantial agreement between domain expert and NLP coding (κ = 0.69). These results suggest a good level of consistency in the interpretation of themes, indicating that the data used at this stage is reliable. Given this reliability, the data is well-suited for further analysis using Method 3, which involves supervised learning.

Table 7. Performance Metrics (Train) for Supervised Learning Method #3
Train Data (N=1425) and Test Data (N=357)

Naïve Bayes									
Theme	**Data Type**	**TP**	**TN**	**FP**	**FN**	**Precision**	**Recall**	**F1 score**	**Accuracy**
Class support	Train Test	62.7% 61.3%	34.6% 30.0%	1.26% 5.60%	1.40% 3.10%	98.0% 91.6%	97.0% 95.2%	96.3% 93.4%	97.3% 91.3%
Interactions	Train Test	13.7% 10.4%	85.2% 86.0%	0.42% 0.30%	0.63% 3.40%	97.0% 97.4%	95.6% 75.5%	96.3% 85.1%	98.9% 96.4%
Problem Solving	Train Test	19.0% 15.4%	78.0% 77.3%	0.63% 0.80%	2.53% 6.40%	96.8% 94.8%	88.3% 70.5%	92.3% 80.9%	96.8% 92.7%
Support Vector Machine (SVM))									
Theme	**Data Type**	**TP**	**TN**	**FP**	**FN**	**Precision**	**Recall**	**F1 score**	**Accuracy**
Class support	Train Test	63.9% 63.0%	34.7% 29.9%	1.12% 5.60%	0.28% 1.40%	98.3% 91.8%	99.6% 97.8%	98.9% 94.7%	98.6% 93.0%
Interactions	Train Test	13.4% 12.3%	85.4% 85.1%	0.28% 0.56%	0.42% 1.40%	98.0% 95.6%	97.1% 90.0%	97.5% 92.6%	99.3% 98.0%
Problem Solving	Train Test	20.6% 17.4%	78.3% 77.0%	0.21% 1.12%	0.98% 4.48%	99.0% 93.9%	95.4% 79.4%	97.2% 86.1%	98.8% 94.4%

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

Supervised Learning, Method 3:

The results of the supervised learning (Naïve Bayes) Method #3 demonstrate high precision, recall, and F1 score for the "Class Support" and "Interactions" themes in the training data, indicating the model's ability to accurately predict these themes ranging from 96.8% to 98.9%. However, the testing data shows a decrease in these metrics (accuracy between 91.3% and 96.8).

The Support Vector Machine (SVM) also exhibits similar trends, with high precision, recall, and F1 score in the training data (ranging from 93.0% to 99.3% accuracy), but a decrease in these metrics in the testing data (93.0% to 98.0%), particularly for the "Interactions" theme. The accuracy in detecting instances of “No response, other, or ambiguous" responses ranged from 98% to 99% for both train and test data.

Discussion

The study conducted at a large public research institution aimed to explore the integration of domain expert interaction into the automatic short answer coding (ASAC) process and the effectiveness of supervised learning techniques in predicting labels for student responses regarding faculty support preferences. The insights gained from both unsupervised and supervised learning methods provide valuable contributions to addressing the research questions posed in this study.

Methodology Research Question (RQ1):

In addressing RQ1, the results from unsupervised learning Method 1 and Method 2 underscore the critical role of domain expert interaction in enhancing the value of automated data analysis in educational research [9]. In Method 1, the NMF-based topic modeling identified four topics, but its accuracy was nuanced [22]. The domain expert's manual coding, based on NMF-generated topics, revealed the model's proficiency in identifying true negatives but highlighted challenges with true positives, especially in the "Affect Support" theme. This difficulty stemmed from instances where responses, algorithmically coded as "Affect Support" by NLP, were appropriately labeled as "No response, other, or ambiguous" with the assistance of domain expert intervention. The domain expert's role proved indispensable in identifying responses that didn't directly address the faculty support experience, a nuance unattainable through NLP alone.

The subsequent Method 2 involved additional preprocessing, guided by domain expert interaction, to eliminate non-answers and ambiguous responses. This refinement resulted in three topics with improved accuracy. The domain expert's coding demonstrated high agreement with NMF in identifying "Problem Solving" and "Interactions" themes but exhibited lower agreement in "Class Support." This outcome emphasizes the refining effect of domain expert interaction in streamlining the ASAC process and focusing the model on pertinent themes.

Cohen's Kappa coefficient results further affirm the value of domain expert interaction, indicating moderate to substantial agreement between two domain expert coders and between domain expert and NLP coding across different themes [21], [24]. This suggests that domain expert involvement significantly enhances the reliability of NLP coding in the context of educational research.

Methodology Research Question (RQ2*):*

In addressing RQ2, the insights generated from Method 2 were instrumental in understanding the types of support students seek from faculty to enhance their learning ("Class Support," "Problem Solving," and "Interactions"). While determining the main themes from the dataset was crucial, equally important was identifying responses that did not answer the faculty support survey questions and did not align with any of the three categories. Therefore, a comprehensive approach included codes assigned to the dataset for the three themes along with responses where domain expert codes were designated as "No response, other, or ambiguous." This dataset was randomly ordered and divided into training and testing sets, which served as input data for the supervised learning method.

The high precision, recall, and F1 scores for "Class Support" and "Interactions" themes in the training data from Method 3 underscore the accurate predictive capabilities of the supervised learning models, Naïve Bayes, and SVM [18], [19]. However, a slight decrease in these metrics in the testing data suggests that while the models are effective, there is room for improvement, particularly in generalizing the models to new, unseen data. The results remained consistent across all themes, ranging in accuracy from the lowest at 91.3% to the highest at 99.3%.

In summary, the study highlights the invaluable contribution of domain expert interaction in managing heterogeneous data, particularly in the context of student responses regarding faculty support [9]. The complexity and diversity of the data, reflecting students varied and indirect impressions, necessitate an approach capable of handling unstructured data effectively. Unsupervised learning methods, especially NMF, emerged as promising for heterogeneous datasets, uncovering latent themes without being confined by predefined categories [16]. The results suggest that integrating unsupervised learning methods with domain expert interaction effectively contributes to the creation of a robust dataset capable of accurately predicting themes in student responses using supervised learning. However, the observed decrease in performance metrics on the testing data highlights the ongoing need for optimization and validation to ensure the models' robustness, generalizability, and applicability in real-world educational research settings.

Implication

The study's findings have several implications for both research and practice. Firstly, the integration of domain expert interaction with unsupervised and supervised learning methods provides a robust framework for analyzing qualitative data in educational research. This approach not only enhances the reliability of automated analysis but also ensures that the resulting dataset is well-suited for predictive modeling using supervised learning techniques. The study's emphasis on the critical role of domain expert interaction in refining themes identified by unsupervised learning methods highlights the importance of human expertise in the data analysis process. This has significant implications for researchers and practitioners, as it underscores the need to combine automated methods with human judgment to ensure the accuracy and relevance of the analysis.

Secondly, the study's results also point to the potential of supervised learning techniques, such as Naïve Bayes and Support Vector Machine (SVM), in predicting themes within student responses. While the models demonstrated high precision, recall, and F1 scores in the training data, further optimization is necessary to ensure their robustness and generalizability to new, unseen data. This suggests that future research and practical applications should focus on refining and validating these models to ensure their effectiveness in real-world educational research settings. Additionally, the study's approach to comparing domain expert and NLP coding results using performance metrics and intercoder reliability assessments provides a valuable method for evaluating the success of different data analysis approaches. This has implications for researchers and practitioners seeking to assess the accuracy and agreement of automated coding methods in comparison to human-coded ground truth data.

In summary, the study's implications highlight the need for a balanced approach that integrates human expertise with automated methods in educational research. By combining unsupervised and supervised learning techniques with domain expert interaction, researchers and practitioners can ensure the accuracy, relevance, and reliability of their data analysis, while also leveraging the predictive capabilities of machine learning models.

Conclusion

The findings of this study underscore the significance of domain expert interaction in enhancing the automated analysis of qualitative data in educational research. Unsupervised learning methods, particularly Non-negative Matrix Factorization (NMF), proved to be effective for handling the complexity and heterogeneity of student responses regarding faculty support. However, the integration of domain expert interaction was essential for refining the themes identified by unsupervised learning. The integration enhanced the reliability of the analysis and helped create a robust dataset, which was then used in supervised learning.

The study also highlights the potential of supervised learning techniques, such as Naïve Bayes and Support Vector Machine (SVM), in predicting themes within student responses. While these models showed high precision, recall, and F1 scores in the training data, further optimization is necessary to ensure their robustness and generalizability to new, unseen data. The observed decrease in performance metrics on the testing data highlights the need for ongoing model refinement to ensure their effectiveness in real-world educational research settings.

In conclusion, the study provides a comprehensive framework for integrating unsupervised and supervised learning methods with domain expert interaction to analyze qualitative data in educational research. The results emphasize the importance of domain expert involvement in refining themes and enhancing the reliability of automated analysis. Future research should focus on optimizing these models and exploring additional methods for integrating expert knowledge to further improve the automated analysis of qualitative data in education.

Limitations

The study is limited by several factors that may have influenced the accuracy and generalizability of the results. Firstly, the potential biases introduced by human annotators during the pre-processing stage could have impacted the outcomes of the analysis. Additionally, the study's narrow focus on a single US research institution and a specific short answer question related to faculty support may limit the generalizability of the findings to other questions or domains. Furthermore, the use of a specific set of natural language processing techniques may restrict the applicability of the results to other techniques or approaches. These limitations should be considered when interpreting the findings and may have implications for the generalizability of the study results.

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] D. Nguyen, M. Liakata, S. DeDeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters, "How We Do Things With Words: Analyzing Text as Social and Cultural Data," *Frontiers in Artificial Intelligence*, vol. 3, p. 62, 2020. [Online]. Available:

https://www.frontiersin.org/articles/10.3389/frai.2020.00062. doi: 10.3389/frai.2020.00062.

[2] 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

[3] L. Fesler, T. Dee, R. Baker, and B. Evans, "Text as Data Methods for Education Research," *Journal of Educational and Behavioral Statistics*, vol. 44, no. 6, pp. 709-732, 2019. doi: 10.1080/19345747.2019.1634168.

[4] Y. Wang and Y. Zhang, "Nonnegative matrix factorization: A comprehensive review," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, June 2012, doi: 10.1109/TKDE.2012.51.

[5] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, “Interactive topic modeling,” *Machine Learning*, vol. 95, no. 3, pp. 423–469, Oct. 2013, doi: 10.1007/s10994-013-5413-0.

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

[7] 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," *Phys. Rev. Phys. Educ. Res.*, vol. 18, no. 1, p. 010141, Jun. 2022. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.18.010141

[8] D. Ramesh and S. K. Sanampudi, "Semantic and Linguistic Based Short Answer Scoring System," *Int J Intell Syst Appl Eng*, vol. 11, no. 3, pp. 246–251, Jul. 2023. [Online]. Available: https://www.ijisae.org/index.php/IJISAE/article/view/3164

[9] 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

[10] 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.

[11] 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.

[12] 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]

[13] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *International Journal of General Systems*, vol. 46, no. 1, pp. 27-36, 2017. doi: 10.1080/03081079.2017.1291635.

[14] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)," *ComTech: Computer, Mathematics and Engineering Applications,* vol. 7, no. 4, pp. 285-294, Dec. 2016. [Online]. Available: https://www.researchgate.net/publication/318963563_Single_Document_Automatic_Text_Summarization_using_Term_Frequency-Inverse_Document_Frequency_TF-IDF

[15] C. Liu, et al., "Research of text classification based on improved TF-IDF algorithm," in *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), IEEE, 2018*. [Online]. Available: https://ieeexplore.ieee.org/document/8492945

[16] Y. Wang and Y. Zhang, "Nonnegative matrix factorization: A comprehensive review," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, Jun. 2012. [Online]. Available: https://ieeexplore.ieee.org/document/8653529

[17] 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].

[18] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," in *Encyclopedia of Bioinformatics and Computational Biology*, pp. 403-412, Elsevier, 2018. doi: 10.1016/b978-0-12-809633-8.20473-1.

[19] [A. Shmilovici, "Support Vector Machines," in *Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Springer*, 2010, pp. 231-247. doi: 10.1007/978-0-387-09823-4_12.

[20] 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.

[21] 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.

[22] 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

[23] 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.

[24] 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