Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #46681

Future-Ready Students: Validating the Use of Natural Language Processing to Analyze Student Reflections on a Remote Learning Group Project

Majd Khalaf, Norwich University

Majd Khalaf recently graduated from Norwich University with a Bachelor's degree in Electrical and Computer Engineering, along with minors in Mathematics and Computer Science. He is passionate about DevOps, embedded systems, and machine learning. Throughout his academic career, Majd contributed to various projects and research in natural language processing (NLP) and computer vision. He served as a Senior AI Researcher at Norwich University's Artificial Intelligence Center, interned at the New England University Transportation Center, and led the IEEE-Eta Kappa Nu Theta Xi chapter as its president. He has completed a Site Reliability Engineering internship at Walmart ASR and is currently a Software Engineer Intern at AtomBeam Technologies.

Toluwani Collins Olukanni, Norwich University

Toluwani Olukanni is a graduate of Norwich University, where he received his bachelor's degree in Electrical and Computer Engineering. His research interests include machine learning and the application of artificial intelligence in medicine, with a focus on computer vision and deep learning. He is passionate about creating AI solutions that enhance human productivity and performance.

Dr. David M. Feinauer P.E., Virginia Military Institute

Dr. Feinauer is a Professor of Electrical and Computer Engineering at Virginia Military Institute. His scholarly work spans a number of areas related to engineering education, including the first-year engineering experience, incorporating innovation and entrepreneurship practice in the engineering classroom, and P-12 engineering outreach. Additionally, he has research experience in the areas of automation and control theory, system identification, machine learning, and energy resilience. He holds a PhD and BS in Electrical Engineering from the University of Kentucky.

Dr. Michael Cross, Norwich University

Michael Cross is Chair and Assistant Professor of Electrical and Computer Engineering at Norwich University teaching classes in the areas of circuits, electronics, energy systems, and engineering design. His research interest is in energy systems, specifically battery electric vehicles and their impact on the electric grid. Cross received degrees from the Rochester Institute of Technology and the University of Vermont.

Ali Al Bataineh, Norwich University

Future-Ready Students: Validating the Use of Natural Language Processing to Analyze Student Reflections

Introduction

First-year Electrical and Computer Engineering (ECE) students from Norwich University and Virginia Military Institute worked remotely on an inter-university team design project. The project was implemented in Spring 2023 and repeated in Spring 2024. At the end of the endeavor, the students completed an end-of-project survey and wrote a reflection about the experience.

Following the initial project offering, the authors employed Natural Language Processing (NLP) techniques to analyze the student reflections. Three unsupervised learning techniques (K-means clustering, Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization) were utilized to identify key themes in the student responses and categorize the topics or themes common among the responses. Preliminary findings based on the Spring 2023 data revealed a set of five common and distinctive themes or topics (performance expectations, collaboration and planning, skill development, problem solving, and evaluation) across the reports from both Institutions and were reported by the authors in a previous publication [1].

Building on this work, the authors repeated the analysis techniques on the data collected during the subsequent project offering. Additionally, the authors used the themes identified from the initial offering to train a classifier. The classifier was used to label and categorize the student reflections from the second cohort based on the themes uncovered or "learned" when analyzing the first cohort of responses.

By replicating the previously reported analyses and using the previous work as the training data set for labeling the results from subsequent project offerings, the authors gained insight into the validity of the technique and the effectiveness of using unsupervised NLP methods to uncover insights from open-ended student responses and reflections. In applying the techniques and validating their efficacy, the authors gained valuable insight into the possibility for NLP and other AI-assisted techniques to be used for academic assessments and for labeling responses as part of qualitative or mixed-methods educational research endeavors.

The methodology section of the paper will detail the application of the techniques to unstructured text collected as free-response student reflections. The findings section presents a comparison of the topics identified among the student responses for two cohorts from the Spring of 2023 and Spring of 2024, focusing on the themes of collaboration and planning (teamwork), as well as problem solving. Lessons learned about the process of applying the techniques, as well as insights gained about the student experience as captured in their reflections, are shared in the conclusions section, along with the authors' recommendations for the use of the AI-assisted process to analyze qualitative data as a means of better understanding the students' project experience.

This work advances the subject of engineering education by showing how automated natural language processing (NLP) techniques may be used to evaluate student reflections, offering a scalable and effective substitute for conventional qualitative analysis methodologies. Tracking

student progress, curriculum improvement, and educational evaluation are all impacted by the capacity to find themes and patterns in student remarks with no manual involvement. Moreover, the results show how AI-assisted reflection analysis may improve feedback loops in the classroom and help teachers make informed decisions about engineering education.

Methodology

The study analyzed feedback from first-year Electrical and Computer Engineering (ECE) students at two academic institutions, Norwich University (NU) and Virginia Military Institute (VMI), over two semesters. The students engaged in a joint project to develop a smart home device, utilizing skills from their introductory courses. Student feedback was collected through a structured reflection exercise conducted after the completion of the project.

In a previous effort, the authors developed a post survey and narrative reflection exercise to capture information about the student experience and determine whether the remote collaboration, multi-university, joint project exercise delivered the desired developmental experiences for the students at the respective universities. The primary research question explored in the previous work involved understanding whether the project construct required the students to work through teamwork challenges that are typically made easier by the highly structured academic day at both institutions. Additional research goals involved identifying key themes in student free responses about the experience related to professional skill development inherent in the project experience. Those efforts revealed that the students felt prepared for the project, but the team formation, as well as general team collaboration and communication, was more challenging than the typical class project. Students also found that their project efforts were more parallelized than typical class exercises.

For the initial study, the authors manually categorized student survey and reflection responses in an attempt to evaluate the project design and better understand the student experience. They also conducted an LDA analysis of student reflections and identified five key themes in the student narrative responses--performance expectations, collaboration and planning, skill development, problem solving, and evaluation.

Following this effort, in the second year offering the joint exercise, the authors applied the methodology described above to answer two key questions:

1. Do the narrative student responses from the second cohort identify the same key themes and discuss the same key topics as the students from the previous cohort?
2. Can the data from the initial project offering be used to train a classifier that could aid in labeling and categorizing new student responses, making the process of extracting meaning from student feedback less time-consuming and less subjective for the instructors?

Data Collection

A written reflection assignment was administered through each school's LMS at the end of the project. A summary of respondent details is shown below in Table 1.

Table 1. Summary of Respondent Details

Question Summary	S23		S24	
	NU	VMI	NU	IB
Total # of students in course	10	11	12	17
# of reflections submitted	8	8	10	10
# Cadets/Civilian	5/5	11/0	5/7	17/0
# Female/Male	0/10	3/8	2/10	1/16

The guided reflection survey was comprised of five sections, as shown below:

1. **Description**
    - What happened during your project experience? (High-level story)
2. **Feelings**
    - How do you feel about the experience? Explain.
3. **Evaluation / Analysis / Conclusion**
    - What behaviors, processes, or skills assisted you in completing this project?
    - What skills do you wish you had developed previously to help you with the project? Why?
    - What did you learn about your partner(s)? How did you learn this?
    - What have you learned about yourself?
    - What have you learned about the engineering process? Why? / Which aspects helped you learn this?
4. **Norming**
    - Did you establish performance expectations and behavior norms? If so, how and when?
    - If something wasn't meeting your expectations, what did you do to correct it?
5. **Action Plan**
    - What advice would you give about how to conduct a joint project like this in the future?
    - What should change?
    - What should be sustained?

The responses from all five survey sections were compiled into combined and institution-specific PDF files for the 2 years of the project.

Preprocessing

The data was pre-processed to prepare it for use in a topic modeling scheme. The preprocessing involved text cleaning (removal of stopwords, special characters, standardization of text case, etc.), tokenization (separation of text into tokens or phrases), lemmatization (reducing words to their base or root form), and vectorization (conversion of phrases into a numerical form) to prepare them for use in a Latent Dirichlet Allocation algorithm [2].

Similarly, preprocessing was also performed on the data to prepare the student responses for use in the selected NLP algorithms and to help those models focus on important aspects of the text. It was discovered that removing special characters, digits, and emoji but leaving stopwords provided the best performance for the classification models, however removing stopwords was an important step for topic modeling. This is rare because stopwords are often removed because they may increase noise and distract the focus from meaningful words; however, in some cases, they preserve the context in sentences.

Topic Modeling

For this study, Latent Dirichlet Allocation (LDA) was used as the main topic modeling technique because it works well with unstructured text and can detect latent themes in narrative student reflections that lack predetermined categories. It is superior to dimensionality reduction methods like PCA (Principal Component Analysis) because of its interpretable output, which organizes words according to co-occurrence patterns. This is especially useful for educational research, where pedagogically relevant topics are crucial. The effectiveness of LDA in earlier educational research has further supported its dependability in examining qualitative thoughts. Furthermore, with little manual involvement, its automation and scalability enable the effective processing of enormous textual datasets. Although other approaches, including K-means clustering and Non-Negative Matrix Factorization (NMF), were taken into consideration, LDA was chosen for engineering education applications due to its theoretical underpinnings, efficacy in small datasets, and ease of interpretation. This choice guarantees an open and insightful examination of the experiences of the students.

The pre-processing of data from the Spring 2023 cohort yielded 511 tokens or phrases. Those phrases were fed into the LDA model, and five prevalent themes or topics were identified. The pre-processing of data from the Spring 2024 cohort yielded 640 tokens or phrases. Those phrases were fed into an LDA algorithm, and five prevalent themes or topics were also identified. The resultant topics were compared and are available in the findings section. It is important to note that although the reflection prompt is provided above as a structured list, the student responses were narrative and far less structured.

Traditional ML Classification

Following the topic modeling approach described above, the authors selected two of the five identified topics – collaboration and planning (teamwork) and problem solving. The authors manually labeled all 511 tokens from the Spring 2023 cohort. Two hundred and eighty-seven (287) of the tokens involved collaboration and planning, 137 involved problem solving, and 384 involved either topic. Those manually labeled tokens served as a training and validation data set for a classifier algorithm.

The labeled tokens from the Spring 2023 dataset were used to create a target label that consisted of only teamwork or problem solving, but not both. This resulted in a binary classification task where the model only had to classify if a sample text was related to teamwork or problem solving. This resulted in 97 tokens involving problem solving and 247 tokens involving teamwork, as shown in Figure 1.

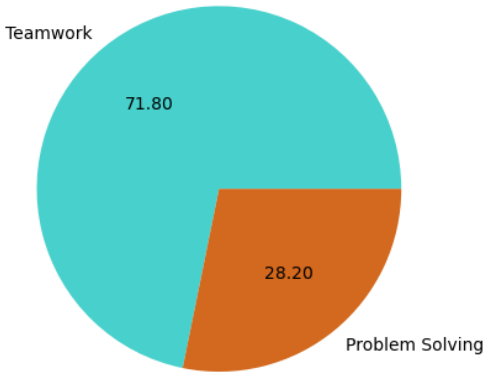

Figure 1. Proportion of each topic in the 2023 cohort following manual labeling.

These 344 total tokens were split into two groups: 80% training and 20% testing using the scikit-learn `train_test_split` method. Due to the apparent imbalance between the two classes of topics, the random oversampling method from `imbalance-learn` was used to balance the distribution only in the training set to allow the model to learn a balanced dataset of examples. After this, both the training and the testing set were vectorized into a numerical format suitable for the machine learning models, using `tfidf_vectorizer` and `casual_tokenize` as the tokenizer [3,4].

Once this was done, an ensemble of 11 ML models was used to fit the training data and make predictions using the testing data for validation. The best performing classifier was then trained on the entirety of Spring 2023 data and used to label and classify the student responses from the Spring 2024 dataset.

DL Classification

A deep learning approach was implemented to classify student reflections, leveraging advanced neural network architectures to enhance the accuracy of thematic analysis. The methodology included importing and pre-processing the dataset, involving tokenization, padding, and word embeddings.

The architecture of the deep learning model was specifically designed to handle narrative data, utilizing a combination of embedding layers and LSTM networks to capture both semantic and sequential patterns in student reflections. The layers and their respective parameters are summarized in Table 2.

Table 2. Summary of Model Architecture and Parameters

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 33, 128)	1,920,000
lstm_2 (LSTM)	(None, 64)	49,408
dropout_2 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 1)	65

The embedding layer, initialized with 128-dimensional vectors, transformed tokens into dense numerical representations, contributing significantly to the model's ability to understand word semantics. The LSTM layer with 64 hidden units captured the sequential dependencies within the text, while the dropout layer with a rate of 30% mitigated overfitting by randomly deactivating neurons during training. Finally, the dense layer served as the output layer, producing a binary classification for each reflection.

This architecture, with a total of 1,969,473 trainable parameters, provided the capacity to model complex narrative structures while maintaining computational efficiency and balancing accuracy. The optimized configuration ensured high accuracy without compromising interpretability, as demonstrated in the findings. A binary cross-entropy loss function with an Adam optimizer was employed. The training data was split into 80% training and 20% testing sets, and the model was trained for 12 epochs with a batch size of 32. Hyperparameter tuning further refined the model to prevent overfitting.

The results of the above methods are presented in the findings section below. A more detailed discussion of the preprocessing and LDA topic modeling technique is available in an earlier publication from the authors [1]. An additional publication from the authors details lessons learned related to the student experience for a project of this sort [5].

Findings

Preliminary findings based on the Spring 2023 data revealed a set of five common and distinctive themes or topics (performance expectations, collaboration and planning, skill development, problem solving, and evaluation) across the reports from both Institutions. The number of topics was restricted to five as a result of using an "elbow method" analysis to determine the point where increasing the number of topics no longer yielded significant performance improvements. The LDA method yielded the top 10 words related to each of the five topics identified. Further interpretation was needed to give a meaningful label or name for each topic. The authors generated labels for each topic based on the grouping of 10 related words, and they also consulted Chat GPT 3.5, requesting a summary category name or label for the words in the spirit of automating the process. Table 3 shows the raw topic words and Chat-GPT generated summative descriptor for the data from both cohorts.

Table 3. Topic words with Chat-GPT descriptors. The two focus topics are shaded in green/blue.

Cohort	Topic Words	Topic Descriptor
23-1	time, norms, performance, feel, expectations, partner, work, learned, did, project	Performance and Expectations
23-2	slides, partners, doing, sent, email, description, got, planning, project, partner	Collaboration and Project Planning
23-3	used, students, developed, skills, worked, process, problem, project, communication, work	Skill Development and Project Work
23-4	issue, work, helped, presentation, needed, think, engineering, partner, learned, project	Problem Solving and Learning
23-5	advice, analysis, feelings, like, work, evaluation, experience, plan, partner, project	Feedback and Evaluation
24-1	like, expectations, feel, meeting, partner, project, work, experience, learn, did	Project Design & Development
24-2	project, sensor, best, tinkercad, code, needed, pump, design, explain, change	Learning & Engineering Process
24-3	learn, aspects, time, project, make, sustained, engineering, helped, process, learned	Reflection & Group Experience
24-4	question, learned, didn't, description, good, happened, reflection, group, experience, project	Project Analysis & Skill Development
24-5	analysis, conclusion, ideas, expectations, learned, wish, developed, time, skills, project	Project Evaluation & Development

Addressing research question 1 regarding the recurrence of themes in both offerings, upon review of the topical themes in Table 1, one should note that elements related to Skill Development (23-3 + 24-4), Reflective Evaluation (23-5 + 24-3), and Project Design and Work Output (23-3 + 24-1) appeared in both datasets along with the two selected themes related to Collaboration and Planning / Group Experience (23-2 + 24-1, 24-3, 24+5, highlighted in green) and Problem Solving / Learning Engineering Process (23-4 + 24-2, highlighted in blue).

The application of topic modeling and the LDA method to the free response reflections from the student project participants helped the authors discover patterns or trends in the responses that may not have been readily apparent. The authors specifically designed this group project to involve remote project work, and the learning outcomes focused less on technical skill development and more on teamwork and project planning in a modern, real-world, remote work environment. Based on the project design and the instructors' intent, it was reassuring that collaboration and project planning, problem solving, and setting of performance expectations were three of the five key topics identified by the automated topic modeling process. It is also worth noting that the process that was developed could scale to analyzing larger volumes of student responses rather quickly—the input was one PDF of student reflections with names removed, and the preprocessing and topic modeling scripts were fully automated. Additionally, the topic modeling analysis was exploratory in that no anticipated topics or labels were provided in advance.

Based on these results, the instructors selected problem solving and teamwork as two key themes, and they manually labeled all of the responses from the Spring 2023 surveys regarding whether a sentence or phrase discussed these themes. It is important to note that the sentiment with which the respondent discussed a theme was not recorded, merely the fact that a topic or theme was present. After manually coding all the responses from Spring 2023, that dataset was used to train a classifier utilizing both deep learning and traditional machine learning techniques to label the responses from Spring 2024. Some of the top five best-performing ML models out of the 11 mentioned in the Methodology section, based on F1 score, were Multinomial Naïve Bayes (MNB), Complement Naïve Bayes (CNB), Voting Classifier (VC), Logistic Regression (LR), and Bagging Classifier (BC), as shown in Figure 2 below.

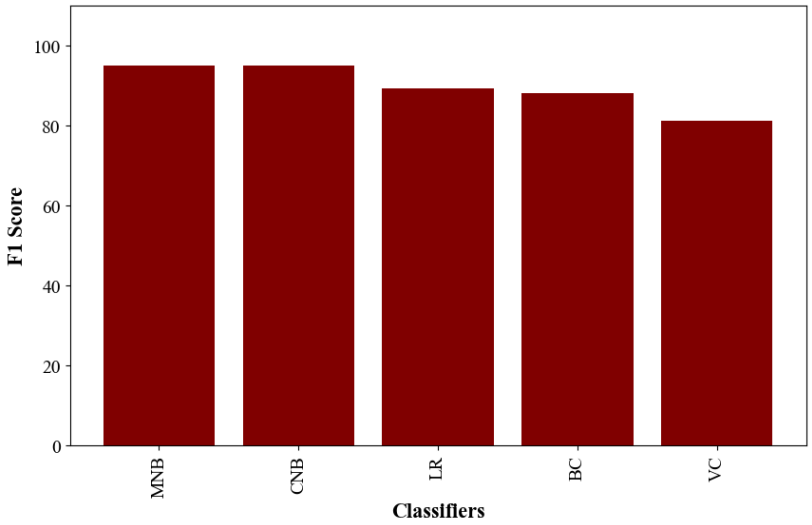

Figure 2. Top 5 classifiers by F1 score.

In response to research 2 regarding the effectiveness of using one year of student data to build a classifier for identifying themes in subsequent project offerings, based on the results from the confusion matrix and the ranking of algorithms above, Multinomial Naïve Bayes was the best-performing model, achieving a testing accuracy of 97.1%. Hence, the Multinomial Naïve Bayes model was retrained using the entirety of the Spring 2023 dataset to classify the Spring 2024 dataset. F1 score refers to the harmonic mean of precision (how many of the predicted positives are positive) and recall (how many of the true positives were correctly predicted). For reference, the F1 score was selected as a performance measure because it balances metrics related to the ability of the algorithm to label true positive matches while also minimizing false positives. F1 score metrics are also helpful with imbalanced datasets such as this [6].

Results from using the classifier to identify phrases or sentiments that address a topic are provided in Table 4. The deep learning model, built using an LSTM architecture, demonstrated significant improvements over traditional ML approaches. With 1,969,473 trainable parameters, the model was trained on the Spring 2023 dataset and tested on the Spring 2024 dataset. The LSTM method achieved a testing accuracy of 98.39%, outperforming traditional ML models in accuracy. Figure 3 illustrates the model's convergence during training, with both training and testing losses stabilizing by the 10th epoch. The use of oversampling methods balanced the

dataset and ensured that the model did not favor one class over the other to enhance its ability to generalize across different datasets [7].

Table 4. Results from classifier application.

Cohort	Collaboration and Planning (N / %)	Problem solving (N / %)	Testing Accuracy
Spring 2023 (Manual, N = 511)	287 / 56.16%	137 / 26.81%	-
Spring 2024 (Auto ML, N = 640)	419 / 65.47%	221 / 34.53%	97.10%
Spring 2024 (Auto DL, N = 640)	398 / 62.18%	242 / 37.81%	98.39%

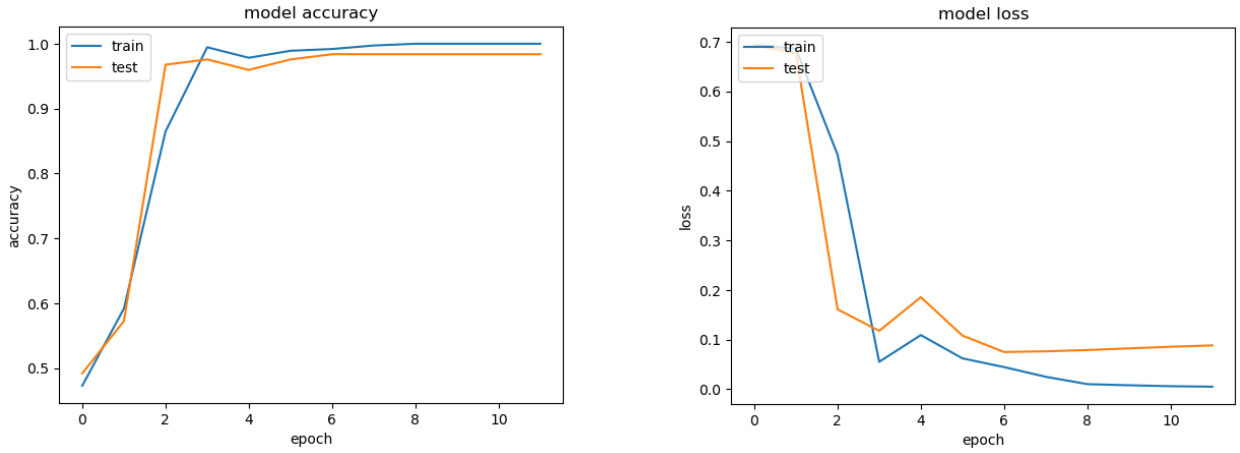

Figure 3. Model accuracy (left) and model loss (right) during training and testing.

The automatic labels for Spring 2024 assume that there is no overlap between both topics, meaning no token can be classified as both teamwork and problem solving. This explains the slight differences in the distribution of teamwork and problem solving from the manually labeled Spring 2023 dataset to the Spring 2024 dataset.

Limitations of the Approach:

This study acknowledges the following limitations:

1. The initial manual labeling of the Spring 2023 dataset introduced subjective bias, as labels were assigned based on instructors' interpretations of student responses. Although efforts were made to ensure consistency, human subjectivity remains an inherent challenge in any manually labeled dataset. Future work could explore crowdsourcing or multiple annotators to enhance reliability.
2. The findings of this study are based on first-year ECE students participating in a specific project-based learning environment. As a result, the insights may not be generalizable to students from other disciplines, higher-level courses, or different educational settings. Extending the analysis to other fields (e.g., mechanical engineering, computer science) could help assess its broader applicability.
3. While LDA provides structured topics, it does not capture semantic nuances or sentence-level meaning as effectively as deep learning models such as BERT-based classifiers. Future

research could explore hybrid approaches, combining LDA for initial topic discovery with deep learning models for finer semantic analysis.

4. The classification approach in this study focused only on teamwork and problem solving, treating them as mutually exclusive categories. However, real-world reflections often contain overlapping themes. Future research could incorporate multi-label classification techniques to address this issue.

Conclusions

The application of Natural Language Processing techniques from the Artificial Intelligence domain was explored for use in analyzing student narrative reflections. The use of an unsupervised topic modeling technique such as LDA helped the authors uncover themes among student survey responses and reflections that might not have been immediately obvious in a way that was automated and scalable. The process was conducted on two different executions of the project and many similar themes emerged from the topic modeling process. This technique allowed the authors to quickly understand the main themes of the student responses and to validate whether the course project was providing the key developmental experiences that the authors intended for their students.

Additionally, AI-assisted classification of student responses was explored. Following the manual labeling of student responses according to an agreed-upon coding manual typical of mixed methods education research, the labeled tokens or phrases were used to train an automatic classifier. The use of the supervised technique allowed the authors to automate the creation of high-quality labeled data for use in further research. This research also served as a learning platform for upper-division students in the ECE program at Norwich University as two students worked for multiple years with the live data set exploring ML and DL topics including topic modeling and classifiers. Working with “known” data that the students had a connection to was meaningful and impactful to their learning and created opportunities for them to extend their understanding beyond typical classroom exercises.

When exploring the results from Table 3, the accuracy of both the ML and DL classifiers shows that the automated technique produced a result consistent with the manual labeling that the instructors would have executed. Additionally, the relative rate of occurrence of the topics (56 vs 65% for Collaboration and Planning and 27 vs 37% for Problem Solving) gives confidence that the classifier and labeling process created was valid and could be more widely applied in future courses and survey analyses.

Future work to improve the performance of the techniques described herein could include the development of a custom list of stopwords to be removed when preprocessing the text or to include engineering education-specific lexicons that could help the LDA model connect certain words. Additionally, although the reflective student responses were unstructured, many of them were linear in their approach to addressing the reflective questions or prompts. Further use of a refined non-negative matrix factorization (NMF) technique could be explored in this case. Lastly, exploration of an ensemble method or non-binary classifier method to allow for labeling or classification of statements that could represent multiple topics could be explored.

Future work with both techniques could improve assessment methods by integrating AI tools to automate the process of interpreting and finding meaning among collected feedback and other educational survey instruments.

References

[1] T. Olukanni, M. Khalaf, M. Cross, D. Feinauer, and A. Al Bataineh, "Full Paper: Future-Ready Students: Survey Analysis Utilizing Natural Language Processing," *Paper presented at 15th Annual ASEE First-Year Engineering Education Conference*, *Northeastern University, MA*, *USA, July 28-30, 2024*, 2024. https://peer.asee.org/4859.

[2] A.-S. Pietsch and S. Lessmann, "Topic modeling for analyzing open-ended survey responses," *Journal of Business Analytics*, vol. 1, Art. no. 2, 2018, doi: https://doi.org/10.1080/2573234X.2019.1590131.

[3] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, Art. no. 1, 2023, doi: https://doi.org/10.3390/info14010054.

[4] Suryaningrum, Kristien Margi, "Comparison of the TF-IDF method with the count vectorizer to classify hate speech," *Engineering, MAthematics and Computer Science Journal (EMACS)*, vol. 5, Art. no. 2, 2023.

[5] D. Feinauer, M. Cross, A. Al Bataineh, T. Olukanni, and M. Khalaf, "Full Paper: Future-Ready Students: Providing Opportunities for Remote Collaboration on an Engineering Design Project," *Paper presented at 15th Annual ASEE First-Year Engineering Education Conference*, *Northeastern University, MA*, *USA, July 28-30, 2024*, 2024. https://peer.asee.org/48599.

[6] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Learning from class-imbalanced data with a model-agnostic framework for machine intelligent diagnosis," *Reliability Engineering & System Safety*, vol. 216, p. 107934, 2021, doi: https://doi.org/10.1016/j.ress.2021.107934.

[7] Z. Wang, S. Kim, and I. Joe, "An improved LSTM-Based failure classification model for financial companies using natural language processing," *Applied Sciences*, vol. 13, Art. no. 13, 2023, doi: https://doi.org/10.3390/app13137884.