# Full Paper: Future-Ready Students: Survey Analysis Utilizing Natural Language Processing

**Toluwani Collins Olukanni, Norwich University**
**Majd Khalaf, Norwich University**

Majd Khalaf is a senior undergraduate student at Norwich University, majoring in Electrical and Computer Engineering. He is deeply passionate about DevOps engineering and machine learning. Majd has contributed to various projects and research in natural language processing (NLP) and computer vision. Currently, he is a Site Reliability Engineering intern at Walmart ASR and a Senior AI Researcher at Norwich University's Artificial Intelligence Center.

**Dr. Michael Cross, Norwich University**

Michael Cross is an Assistant Professor of Electrical and Computer Engineering teaching classes in the areas of circuits, electronics, energy systems, and engineering design. Cross received degrees from the Rochester Institute of Technology and the University of Vermont.

**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.

**Ali Al Bataineh, Norwich University**

# Full Paper: Future-Ready Students – Survey Analysis Utilizing Natural Language Processing


## Abstract

First-year Electrical and Computer Engineering students from two institutions engaged in a collaborative project to develop a smart home device using sensors and actuators learned in their introductory courses. They reflected on the project, and their feedback was analyzed using unsupervised and Natural Language Processing techniques like K-means clustering and Latent Dirichlet Allocation. Key methods included data preprocessing and cleaning. AI tools like TF-IDF vectorization and ChatGPT helped identify key themes such as "PROJECT," "PARTNER," "WORK," and "LEARNED." This study highlights NLP's role in enhancing educational strategies and understanding student experiences.


## Introduction

The application of Natural Language Processing (NLP) approaches in the classroom has the power to transform instruction and improve student performance. With the increasing amount of textual data generated by student feedback, assignments, and reflections at educational institutions, NLP can offer a more profound understanding of students' learning processes and experiences [1] [2]. This study used unsupervised NLP to analyze student input from two distinct academic institutions to identify trends and insights that may guide future curricular modifications.

The main contributions of this work are twofold: first, it presents a methodological framework for analyzing educational data using unsupervised NLP techniques that can be applied in similar settings; second, it identifies key topics in student feedback that are pivotal for curricular adjustments and enhanced teaching approaches.

## Methodology

### *Data Collection and Preprocessing*

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). 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. Sixteen students completed the assignment, 8 from NU and 8 from VMI. The guided survey was comprised of 5 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, and the raw data from the student reflections were read and processed using the PyPDF2 library [3] for reading and extracting text, Pandas library [4] for data manipulation, and the Regular Expressions library [5] for text cleaning. The preprocessing steps included [6]:

- Normalization: Converting text to a consistent format (e.g., lowercasing).
- Tokenization: Splitting text into individual words or tokens.
- Removing stop words and punctuation: Eliminating common words and punctuation that do not contribute to the analysis.
- Lemmatization: Reducing words to their base or dictionary form.

The preprocessed data from the reflection surveys were then broken down into approximately 600 phrases. This detailed pre-processing step was essential for preparing the data for algorithmic analysis, which included using the TF-IDF vectorizer [7] to convert text into numerical vectors, reflecting the uniqueness of words by comparing their document frequency.

**Methods**

The cleaned and processed data, which included responses to all five survey sections, were then analyzed using two unsupervised NLP techniques:

1. Clustering with K-means:
    - This method grouped similar data points based on shared characteristics. The K-means algorithm was specifically used to partition the feedback into predefined clusters by measuring the Euclidean distance from each point to the cluster's centroid. This technique helped in identifying and grouping analogous topics, uncovering patterns and commonalities across the text data [8] [9] [10].
2. Topic Modeling with Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NNMF):
    - LDA is a probabilistic model that assumes documents are mixtures of topics, where each topic is characterized by a distribution of words. This model was used to assign each document a probability of belonging to a given topic, thus facilitating a deeper understanding of the text's thematic composition [11] [12].
    - NNMF decomposes high-dimensional text data into lower-dimensional matrices, representing latent topics. It emphasizes the extraction of additive, non-negative features of the text, enhancing interpretability [13].

The application of these techniques allowed for a detailed mapping of the thematic landscape of the datasets, providing insights into the prevalent topics and aiding in the effective organization of the content for further qualitative analysis.

**Results and Discussion**

Through analysis, five main topics across the surveys from NU and VMI were identified. The Elbow Method was utilized to determine the optimal number of clusters. This method helped identify the point where increasing the number of clusters no longer provided significant improvement in variance, thus indicating the best number of clusters for a K-means analysis. Initially, the number of topics was arbitrarily set to five for the LDA analysis. Further experiments with different numbers of topics were conducted to test the clarity and distinction of the topics. It was observed that increasing the number of topics beyond five resulted in less distinct and more overlapping themes, which diminished the analytical value of the topics. Ultimately, the LDA method was chosen as the preferred approach as it provided clearer group distinctions.

The top ten words from each topic for the combined, NU and VMI surveys were determined and are shown below:

Combined surveys:

- Top 10 words for topic #0: ['time', 'norms', 'performance', 'feel', 'expectations', 'partner', 'work', 'learned', 'did', 'project']
- Top 10 words for topic #1: ['slides', 'partners', 'doing', 'sent', 'email', 'description', 'got', 'planning', 'project', 'partner']
- Top 10 words for topic #2: ['used', 'students', 'developed', 'skills', 'worked', 'process', 'problem', 'project', 'communication', 'work']
- Top 10 words for topic #3: ['issue', 'work', 'helped', 'presentation', 'needed', 'think', 'engineering', 'partner', 'learned', 'project']
- Top 10 words for topic #4: ['advice', 'analysis', 'feelings', 'like', 'work', 'evaluation', 'experience', 'plan', 'partner', 'project']

NU surveys:

- Top 10 words for topic #0: ['like', 'behavior', 'meeting', 'solution', 'performance', 'norms', 'expectations', 'partner', 'project', 'work']
- Top 10 words for topic #1: ['ideas', 'like', 'team', 'partner', 'projects', 'communication', 'time', 'skills', 'learned', 'project']
- Top 10 words for topic #2: ['able', 'arduino', 'needed', 'learned', 'working', 'good', 'wish', 'better', 'partner', 'project']
- Top 10 words for topic #3: ['goal', 'smaller', 'tasks', 'helped', 'communication', 'process', 'make', 'decision', 'work', 'project']
- Top 10 words for topic #4: ['need', 'used', 'challenges', 'people', 'able', 'learned', 'think', 'communication', 'work', 'project']

VMI surveys:

- Top 10 words for topic #0: ['little', 'helped', 'did', 'making', 'email', 'time', 'lot', 'ideas', 'partner', 'project']

- `Top 10 words for topic #1: ['design', 'work', 'better', 'did', 'skills', 'engineering', 'presentation', 'process', 'learned', 'project']`
- `Top 10 words for topic #2: ['work', 'day', 'presentation', 'lack', 'expectations', 'process', 'like', 'did', 'project', 'partner']`
- `Top 10 words for topic #3: ['learned', 'change', 'meet', 'communicating', 'really', 'didn', 'time', 'work', 'partner', 'project']`
- `Top 10 words for topic #4: ['communication', 'work', 'complete', 'learned', 'code', 'think', 'know', 'experience', 'circuit', 'project']`

After this, those topics were given meaning names in two different ways. Based on the information provided above, the first approach was to ask ChatGPT [14] to provide some summary suggestions. This was confirmed through a manual categorization or labeling of the topics by one of the instructors. The results of these techniques for the combined and individual institution datasets are shown below in Table 1.

Table 1: Topic summary for survey datasets utilizing vectorizing and LDA.

| Topic # | Manual: Combined | GPT-3.5: Combined | GPT-3.5: NU | GPT-3.5: VMI |
|---|---|---|---|---|
| 0 | Beneficial | Performance and Expectations | Teamwork and Collaboration | Collaboration and Support |
| 1 | Partners | Collaboration and Project Planning | Innovative Ideas and Project Management | Design and Skill Development |
| 2 | Outcomes | Skills Development and Project Work | Skill Development and Improvement | Meeting Expectations and Challenges |
| 3 | Reflection | Problem-solving and Learning | Goal Setting and Task Management | Learning and Effective Communication |
| 4 | Improvements | Feedback and Evaluation | Adaptability and Overcoming Challenges | Communication and Technical Skills |

After identifying the topics, they were reassigned to the phrases based on probability, establishing their distribution across the topics, as shown in Figure 1. This analysis provided insights into the prevalence and associations of specific phrases with each topic to enhance the understanding of the dataset.

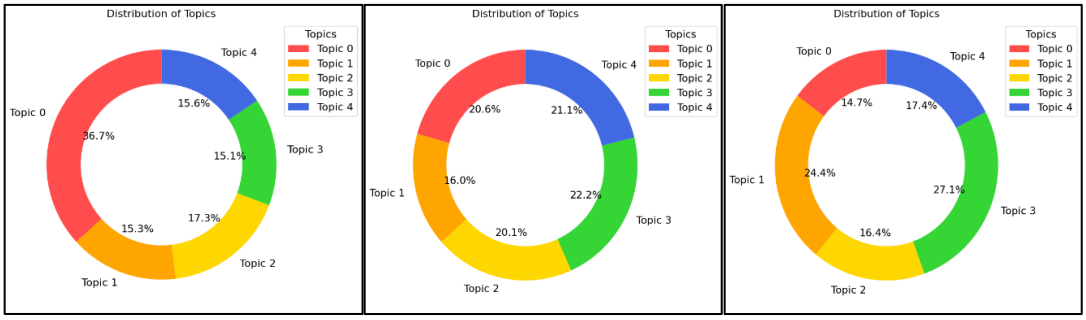


Figure 1: Distribution of topics in the combined (left), NU (center), and VMI (right) datasets.

The text extracted from each dataset (combined, NU, and VMI) was used to produce word clouds that display the most recurring words from the student responses [15] [16]. The word clouds generated from the survey data are shown in Figure 2.


Figure 2: Combined (left), NU (center), and VMI (right) word clouds.

Across all three-word clouds, there are four words that stand out from the rest: "PROJECT," "PARTNER," "WORK," and "LEARNED." In comparing the word clouds of Figure 2 to the response themes summarized in Table 1, one may notice that the AI summaries of the LDA are less superficial than the word clouds. The words "PROJECT," "PARTNER," "WORK," and "LEARNED" would presumably be prevalent in any reflection on a group class project and are also likely prevalent due to the reflection prompt. However, the themes identified in Table 1 provide more insight into the experience's meaning for the students and support inquiry into the common and unique attributes of the experience for students at NU and VMI.

**Challenges**

Use of NLP techniques such as Topic Modeling comes with several limitations. An explanation of those challenges and the efforts taken to mitigate their impact are detailed below.

- Data Quality and Preprocessing: Feeding a huge amount of unorganized data containing grammatical errors, slang, and abbreviations in students' responses can negatively affect the performance of NLP models. This can be mitigated by proper and thorough preprocessing such as stemming, lemmatization, and tokenization.
- Interpreting the model: Topic Modeling algorithms such as LDA and NNMF are difficult to interpret and understand. Using visualization tools such as word clouds or donut plots help to understand the various topics, their distribution, and central ideas.
- Bias in Data and Models: Some data contain inherent biases and can skew the model results if not considered. Validating the categories of topics using diverse opinions, for example, human insight and AI insight, can help to minimize the impact of the biases on the results, making the output more representative of the case study.

**Conclusion**

This study has demonstrated the powerful role of unsupervised NLP in the educational assessment of student project reflections. Through meticulous analysis utilizing methods such as K-means clustering and Latent Dirichlet Allocation, distinct and overlapping themes were identified across the reports from Norwich University and Virginia Military Institute. The findings underline the potential of NLP tools to enhance educational strategies by providing deeper insights into student experiences and learning outcomes. Future work can expand on these techniques to explore additional educational settings, aiming to further enrich student learning and pedagogical approaches. This research contributes to the academic field by improving assessment methods and paves the way for integrating more AI-driven tools into educational evaluations.

Lastly, if the reader is interested in additional details on the execution of and student experience with the class project that served as the test case for the NLP analysis techniques detailed in this paper, the authors have a separate paper published at this same conference (FYEE 2024) that provides those details.

**References**


[1] Y. Lan *et al.*, "Survey of Natural Language Processing for Education: Taxonomy, Systematic Review, and Future Trends," *arXiv (Cornell University)*, Jan. 2024, doi: https://doi.org/10.48550/arxiv.2401.07518.

[2] T. Shaik *et al.*, "A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis," *IEEE Access*, vol. 10, pp. 56720–56739, 2022, doi: https://doi.org/10.1109/access.2022.3177752.

[3] A. Sumich and E. Melnikov, "PDF with Python - Read, Generate, Edit, and Extract Text with Our Examples," *diveintopython.org*. https://diveintopython.org/learn/file-handling/pdf (accessed May 09, 2024).

[4] W. McKinney, "User Guide — pandas 2.2.2 documentation," *pandas.pydata.org*. https://pandas.pydata.org/docs/user_guide/#user-guide (accessed May 12, 2024).

[5] U. Malik, "Using Regex for Text Manipulation in Python," *Stack Abuse*, Aug. 17, 2018. https://stackabuse.com/using-regex-for-text-manipulation-in-python/ (accessed May 12, 2024).

[6] V. H. T. Duong, "NLP Text Preprocessing: Steps, tools, and examples," *Medium*, Oct. 21, 2020. https://towardsdatascience.com/nlp-text-preprocessing-steps-tools-and-examples-94c91ce5d30

[7] Codersarts Ai, "Different techniques for Text Vectorization.," *Codersarts AI*, Jan. 27, 2021. https://www.ai.codersarts.com/post/different-techniques-for-text-vectorization (accessed May 09, 2024).

[8] C. Sampaio, "Definitive Guide to K-Means Clustering with Scikit-Learn," *Stack Abuse*, Jul. 07, 2022. https://stackabuse.com/k-means-clustering-with-scikit-learn/

[9] K. Arvai, "K-Means Clustering in Python: A Practical Guide – Real Python," *realpython.com*, 2020. https://realpython.com/k-means-clustering-python/

[10] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means Clustering Algorithms: a Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data," *Information Sciences*, vol. 622, no. 622, Dec. 2022, doi: https://doi.org/10.1016/j.ins.2022.11.139.

[11] K. Pykes, "What is Topic Modeling? An Introduction With Examples," *DataCamp*, Oct. 2023. https://www.datacamp.com/tutorial/what-is-topic-modeling

[12] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *Journal of Big Data*, vol. 6, no. 1, Oct. 2019, doi: https://doi.org/10.1186/s40537-019-0255-7.

[13] P. Kherwa and P. Bansal, "Topic Modeling: A Comprehensive Review," *ICST Transactions on Scalable Information Systems*, vol. 0, no. 0, p. 159623, Jul. 2018, doi: https://doi.org/10.4108/eai.13-7-2018.159623.

[14] OpenAI, "ChatGPT-3.5," *openai.com*. https://openai.com/chatgpt/

[15] T. Plagata, "How to Create Beautiful Word Clouds in Python - Towards Data Science," *Medium*, Jan. 28, 2021. https://towardsdatascience.com/how-to-create-beautiful-word-clouds-in-python-cfcf85141214

[16] D. Vu, "Python Word Clouds Tutorial: How to Create a Word Cloud," *www.datacamp.com*, Feb. 2023. https://www.datacamp.com/tutorial/wordcloud-python