Analysis of Vlog Transcripts using Topic Modeling, Summarizer, and Cluster Analysis

Upmanyu Singh[1], Tiffany-Rose Sikorski[2], Erica Wortham[3], and Ekundayo Shittu[4]

[1]Columbian College of Arts and Sciences, The George Washington University, Washington, D.C.

[2]Graduate School of Education and Human Development, The George Washington University, Washington, D.C.

[3]GW Innovation Center, School of Engineering and Applied Science, The George Washington University, Washington, D.C.

[4]Engineering Management & Systems Engineering, The George Washington University, Washington, D.C.

2023 IEEE Frontiers in Education Conference (FIE) | 979-8-3503-3642-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/FIE58773.2023.10343205

***Abstract*—This paper demonstrates the use of natural language processing (NLP) models to analyze qualitative data for engineering education research purposes. Three NLP models are applied: topic modelling, which identifies salient keywords in text; summarizer, which extracts and concatenates sentences with unique meaning from a text; and cluster analysis, which groups texts together based on similar word sequences. The study applied these techniques to short video logs, or vlogs, collected as part of a case study of undergraduate engineering students' exposure to a social innovation curriculum. The curriculum aimed to teach students to use their engineering skills to identify and address issues that cause suffering for marginalized communities. In the vlogs, students responded to prompts asking them to describe what social issues they are passionate about, how they relate to these issues, and how they would propose exploring and addressing those issues. Vlogs were transcribed, pre-processed, and then examined with each technique to identify patterns within and across the prompts. Widely-recognized limitations of NLP techniques included the potential loss of important contextual information and the dependence on large volumes of data to produce valid and reliable results. Despite these limitations, the combination of three techniques was effective for locating high priority transcripts within the data corpus, identifying themes within and across vlogs, and supporting longitudinal analysis of student responses. Previous literature has documented the utility of topic modeling and other NLP techniques to analyze large volumes of written text on, for example, course evaluations or student writing assignments. Importantly, this study demonstrates the novel and meaningful application of topic modelling, summarizer, and cluster analysis to analyze a relatively small corpus of transcript data. Given these results, we are optimistic about the potential for NLP approaches to complement other analysis techniques and make the analysis of transcript data more efficient and feasible.**

***Keywords*: Natural Language Processing, topic modeling, cluster analysis, social innovation**

I. Introduction

Transcript data, such as those collected during interviews, focus groups, and classroom recordings, are an important source of data for many engineering education research projects. But analyzing these transcripts has been a significant challenge for researchers due to the time-consuming and labor intensive nature of manual analysis [1]. Thus, there is interest in exploring how automated techniques can complement, or even replace, traditional manual approaches. [2].

In this study, we employed three automated techniques – topic modelling [3], summarizer, and cluster analysis [4] to analyze short transcripts generated in an engineering education research project. Topic modeling is a powerful statistical method that identifies latent topics in a collection of texts, enabling researchers to uncover underlying themes and patterns without reading each transcript manually [3]. Summarizer, a text summarization tool, can extract the most important sentences from a given text and provide concise summaries of the identified topics [5]. Cluster analysis, on the other hand, groups similar transcripts based on the identified topics, offering a better understanding of the data and revealing patterns across different transcripts [4].

We draw on the work of previous studies that have demonstrated the effectiveness of these techniques in various contexts. Our goal is to demonstrate the potential of these methods in handling a relatively small data set comprised of transcripts from students' self-recorded video logs, or vlogs.

II. Literature Review

Natural Language Processing (NLP) has been gaining attention for its potential applications in education, particularly in the field of engineering. A GoogleScholar search of "natural language processing" AND "engineering education" yields approximately 7000 results, underscoring the growing interest in this area. Many of these studies explore how NLP can assist with the analysis of qualitative data, which can require substantial time and expertise to complete manually [6].

An exhaustive review of NLP applications is beyond the scope of this paper, but among the notable studies is one conducted by Bhaduri and colleagues by [7], which

utilized NLP to analyze open-ended student survey data collected during an online course during the COVID-19 health emergency. The study, similar to ours, employed TF-IDF (Term Frequency-Inverse Document Frequency) vectorization [8], a common technique used to identify key words across a text corpus. The research team combined TF-IDF with sentiment analysis to understand how students felt about course assignments, to inform the redesign of those assignments in the future. This aligns closely with our own work, where we also found NLP techniques useful for evaluating the quality of vlog assignments.

Another noteworthy work by Ganesh and colleagues [9] applied NLP in assessing student learning outcomes in online engineering courses. They introduced a task known as response construct tagging (RCT) that tags student responses to survey questions with six constructs related to transformative experiences and engineering identity. Utilizing RoBERTa, a state-of-the-art language representation model, they classified student responses, achieving a good accuracy. The study concludes with a comprehensive qualitative analysis of the model's performance, showcasing the significant potential of NLP in facilitating large-scale pedagogical research in engineering education. This study shares similarities with our research, in which we employed Natural Language Processing methods to evaluate the quality of vlog assignments using transcript data.

Arbogast and Montfort [10] applied NLP techniques to analyze interview transcripts collected from engineers at multiple points during the early years of their careers. The transcripts had been gathered as part of another research project, so the application of NLP techniques was a secondary analysis of that data. The authors found that, despite the limited data set, they could make claims about changes in engineers conceptual reasoning over time via NLP techniques. Similarly, in our study, we applied NLP techniques to analyze transcript data collected from engineering students at multiple points during an academic semester. Inspired by work in other domains [11], we explored the potential of combining multiple NLP techniques to analyze the transcript data.

III. Methodology

A. Data Corpus

The data for this study was gathered over two iterations of a course sequence designed to equip engineering students with the knowledge and skills to address complex societal challenges. The "Social Innovation in Engineering" course sequence consisted of two 15-week, 1-credit seminars. In Seminar I, students engaged in a collaborative, case-based learning environment fostering curiosity, tinkering, and critical thinking. They were introduced to a methodological toolkit for problem-solving, incorporating systems thinking, human-centered design, and ethnographic field research. Students were exposed to social innovation as a concept across several general arenas – from government and industry to non-governmental and advocacy domains – and are invited to imagine themselves social innovator-engineers. This program emanated from a funded workshop on ways to deepenm the engagement of engineering students in their studies [12].

In Seminar II, students built upon the foundational knowledge and skills acquired in the first seminar and delve deeper into concepts and methodologies while tackling real-world social innovation challenges. This course followed the iterative steps of human-centered design beginning with a period of open inspiration and discovery in which students were given the space to think about problems they would like to solve. Over the course of the term, students were guided to deeply understand the problem from an empathetic perspective, bringing people or communities who are directly experience the problem into the design process. Students worked with local community-based partners such as conservation agencies, after school education programs, homeless serving charities to tackle issues of clean water management and green, urban infrastructure, equity in education and housing respectively.

Multiple measures were used to assess the impact of the two courses on students' understanding of and commitment to social innovation, including surveys, video logs, copies of ungraded class assignments and projects. The NLP techniques (topic modelling, summarizer, and cluster) were applied to one of these measures – the video logs, or vlogs. Vlogs consisted of brief, typically 2–5 minute self-recorded responses that students submitted as homework assignments. Students were provided prompts for each vlog. Each vlog generated one transcript for automated analysis. Vlogs were transcribed manually. (If students provided a transcript with their recording, the transcript was checked and edited manually.). Students submitted a total of 173 vlogs. The analysis which is present in this paper is of 75 transcripts collected from the first cohort of students who completed Social Innovation I. However, there were 36 vlogs for same course next year but for the fewer students. For 2021 Social Innovation 1 total of 10 students participated and it got reduced to total of 5 students in 2022. So, we did not use those vlogs for analysis because there should be equal amout of data on both side for analysing. These transcripts included in the automated analysis had an average word count of 550 words and a total word count of 42700. Please note these numbers are before preprocessing, so therefore they include all the stopwords which we excluded before applying all the NLP models,

as described in the next section.

B. NLP Models Used

We utilized topic modelling, summarizer, and cluster analysis to analyze the vlogs. But before using any NLP modelling, a few basic pre-processing steps need to be completed to prepare the vlog transcripts for analysis. First, word tokenization [13] breaks down text into individual words or tokens. For example, one vlog read, "I definitely want to expand my cultural awareness as well as um my educational awareness" Through word tokenization, this becomes ["I", "definitely", "want", "to",..."awareness"]. Next, NLTK [13] stopword removal eliminates common words that would disproportionately impact the results, and unnecessarily increase processing time, like "the" and "is". The stopword list is extended which includes vocal disfluencies like "um" and "uh". In the previous example, the sentence would become ["definitely," "want," "expand" "cultural", "awareness", "well", "educational", "awareness"]. These stopwords extend from the gist link. This link has the extended stopwords list (for example: "um","uh," etc.). Following this, bigrams (pairs of consecutive words) and trigrams (groups of three consecutive words) are created to capture important sequential information; e.g., bigrams like ["definitely want", "definitely expand", "definitely cultural"...and so on] and trigrams like ["definitely want expand", "want expand cultural"...and so on]. Finally, lemmatization reduces words to their base or root form, standardizing different forms of the same word, e.g., ["cultural" to "culture"].

After pre-processing [13], we generate a bag-of-words representation, where each vlog transcript is represented by a vector in a high-dimensional space (dimension equal to the number of unique words in all documents). Each dimension corresponds to a unique word, and the value in each dimension corresponds to the frequency of that word in the transcript. This is mathematically represented as $d = [w_1, w_2, \ldots, w_n]$, where d is the document and w_i is the frequency of word i in the transcript.

At this point, we can begin topic modelling. Topic modelling is used for unsupervised (not labelled) classification of documents, similar to clustering on numeric data, which finds natural groups of words (topics) even when we do not know exactly what we are looking for. In the model, the words in the corpus are examined individually and concurrently, as well as the relationship between each word. Additionally, the frequency of words that co-occur in the corpus is determined. Latent Dirichlet Allocation (LDA) is the model used for topic modelling [14]. LDA is applied for dealing with massive amounts of data and locating relevant keywords within them [14]. Although many topic modelling models have been developed. Specifically for finding the topic within the transcript data, LDA mallet is the model which is used [3]. This model is superior to the classic LDA model in terms of accuracy [15]. The mallet uses gibbs sampling, which is computationally slow but produces more accurate results [3]. On the other hand, classic LDA is fast to compute and it employs the variational Bayes technique, but it is less accurate. This is mathematically represented as

$$P(w|d) = \sum_{t \in T} P(w|t)P(t|d) \tag{1}$$

where $P(w|d)$ is the probability of word w in transcript d, T is the set of all topics, $P(w|t)$ is the probability of word w given topic t, and $P(t|d)$ is the probability of topic t in transcript d.

Finally, we calculate the coherence score, a measure of the quality of the learned topics [16]. A high coherence score indicates that the words in a topic are semantically close to each other. This can be mathematically represented by a variety of measures, such as

$$C = \frac{1}{|W|} \sum_{w \in W} \log \frac{P(w|W)}{P(w)} \tag{2}$$

where C is the coherence score, W is the set of all words in a topic, and $P(w|W)$ and $P(w)$ are the probabilities of word w given the set of words W and individually, respectively.

After topic modelling, we employed text summarization [13]. Text is summarized into 3-4 sentences, enabling researchers to grasp the context of the vlog by simply reading these sentences [13]. Whereas topic modelling involved breaking the transcript into words using the tokenzer, the implementation of the summarizer involves breaking the vlog transcript into different sentences using sentence tokenizer [13]. Then each sentence is converted into a sentence vector. The similarity between pairs of sentence vectors is calculated as

$$\text{cosine similarity} = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{|\boldsymbol{A}||\boldsymbol{B}|} \tag{3}$$

Based on the sentence similarity, a sentence matrix is constructed, and the PageRank [17] algorithm is applied to rank the sentences by importance. The PageRank $PR(p)$ of a page p is given by

$$PR(p) = (1 - d) + d \sum_{q \in M(p)} \frac{PR(q)}{L(q)} \tag{4}$$

where $M(p)$ is the set of pages that link to p, $L(q)$ is the number of outbound links on page q, and d is a damping factor (usually set to 0.85). PageRank, under the hood, take the sentence matrix, which contains all the cosine similarity score for each of the sentences with each other.

For example, if a vlog has 5 sentences, the sentence matrix will be 5x5 matrix showing how similar each sentence is to the other five. Pairs of sentences which are very similar are presumed to be repeating, and thus only one in the pair will be retained by the PageRank [17] algorithm to generate the summary. When a pair of sentences are not similar, PageRank [17] will keep both for the summary. The resulting summary includes all of the sentences from the transcript that convey unique meaning relative to the others [18], [17]. An embedded assumption of the PagRank [17] approach is that all of the sentences within the transcript are relevant to the overall transcript; for example, a sentence which is "off-topic" would be very different from the other sentences, and retained in the summary, even if it was not central to the point the student was trying to make in their vlog.

After analyzing the transcripts using LDA and summarization techniques, we further explored the data by employing cluster analysis. The pre-processing steps are the same as for the topic modelling process. After pre-processing, TF-IDF (Term Frequency-Inverse Document Frequency) transformation was applied to the processed data. TF-IDF is a statistical measure used to evaluate the importance of a word to a document in a corpus. The TF-IDF value is calculated as:

$$TF\_IDF_{i,j} = TF_{i,j} \times log\left(\frac{N}{DF_i}\right) \tag{5}$$

where $TF_{i,j}$ is the number of occurrences of word i in transcript j, DF_i is the number of transcripts containing the word i, and N is the total number of transcripts. The resulting TF-IDF vectors are then used as input to the K-Means++ algorithm for clustering [4]. The K-Means++ algorithm aims to minimize the within-cluster sum of squares (WCSS), given by the equation:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{6}$$

where k is the number of clusters, C_i is the set of observations that belong to cluster i, and μ_i is the centroid of cluster i.

IV. Results

Each of the techniques provided valuable information to analyze within a single vlogs, and compare results across two or more vlogs.

Different representations of the results of the topic modelling anlaysis are shown in Tables I, II, and III. By reviewing the top keywords for each vlog, those with a probability greater than 0.08, we were able to determine which prompts elicited the most student commentary about social innovation. For example, vlog 5 asked students to describe what problem they would solve if they had a "policy magic wand." As shown in Table I, the topics and keywords extracted from students' responses to vlog 5 (access, community, etc.) are directly relevant to the core concept of social innovation and the amelioration of suffering of marginalized groups. This suggests that the vlog 5 prompt elicited useful data for answering our research questions. In contrast, vlog 10 asked students to describe a moment of failure. The topics and keywords shown in Table II (moment, work, hard, year, school, etc.) indicate that students described a moment of *academic* failure, rather than a moment of failure associated with a social innovation project. By reviewing the topic modelling results for each vlog in this way, we were able to identify prompts that yielded high quality information about student understanding of social innovation, and which vlogs did not elicit such evidence, generating Table III. This information is useful for deciding where to focus our attention for manual analysis of the data, and also which vlog prompts we would recommend for future use in social innovation

TABLE I
Topic and Keywords for Vlog 5

Topic ID	Word	Word Importance	Word Count
0	access	0.156	21
0	community	0.104	15
1	homeless	0.119	12
1	build	0.089	9
2	thing	0.140	17
5	basically	0.081	10
6	policy	0.245	35
6	people	0.126	24
7	kind	0.116	16

TABLE II
Topic and Keywords for Vlog 10

Topic ID	Word	Word Importance	Word Count
0	moment	0.134	11
0	work	0.122	12
0	hard	0.085	7
1	year	0.129	17
1	school	0.106	14
2	time	0.121	16
2	people	0.110	11
4	failure	0.208	25
4	kind	0.178	18
5	class	0.137	10
5	thing	0.082	6
6	learn	0.198	17
6	lot	0.081	7

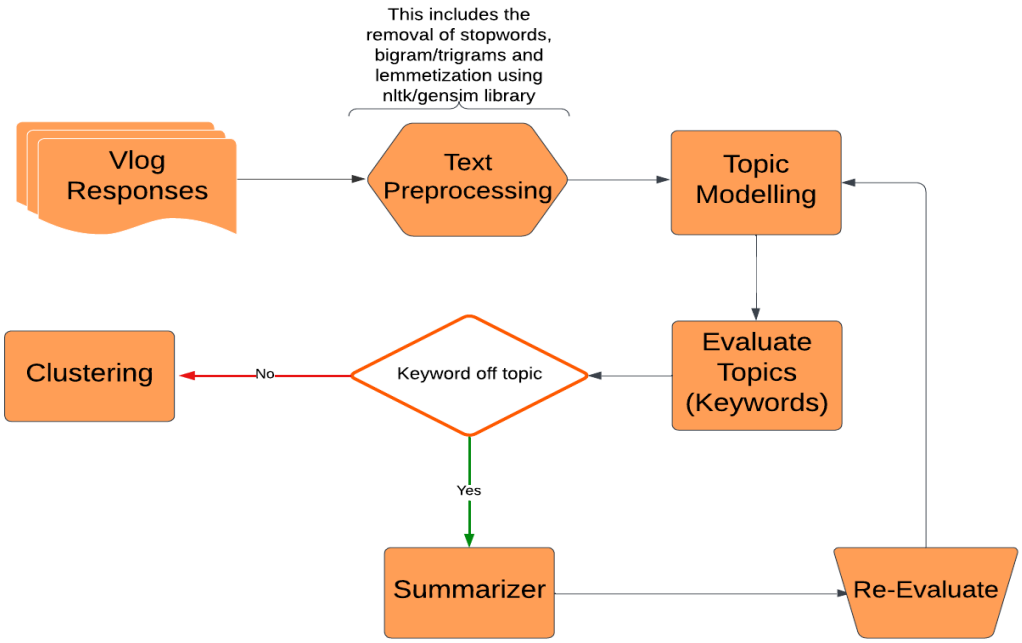

Fig. 1. Outline of the Process

teaching and research. For example, we might revise the prompt for vlog 10 in the future to ask students to describe a moment of failure associated with a civic engagement or community-oriented project. Sometimes, the topic modeling results revealed some seemingly "out-of-context" words which warranted further investigation. For example, vlog 6 asked students to describe their relationship to and role in a social problem. As shown in Table III, one of the keywords extracted via topic modelling is "kid." It is not immediately obvious how this keyword is related to related to the theme of the prompt or the topic of social innovation. To check whether "kid" is used in a way that is related to social innovation, we could manually review each vlog to see the word used in context. For a small number of words, this would not take much time, but for a larger set of words or transcripts, we were able to make this process more efficient by implementing the summarizer tool. A portion of the summarizer's output for the vlog mentioning "kid" is as follows:

so luckily they were able to find really good parents, and they kind of had a happy ending and they could definitely grow out of their behavioral problems because they were so young, but with older kids like this is definitely more difficult. when i was working with the toddlers..

The words "kid" is used in the context of the student discussing a social issue – the foster care system. Thus, the keyword "kid" is in fact relevant to the social innovation course and the prompt. A limitation of topic modeling is that words are taken out of context, which can change or remove layers of meaning. Since we relied on these decontextualized keywords to determine the priority of the prompts as data sources, the summarizer provided an important way to confirm or double check our rankings for each prompt.

After obtaining the summarized text from the vlogs, K-means++ [4] clustering is employed to group students' responses based on their responses (Table IV). The analysis generated 12 clusters, and a subset of the results are depicted in Table IV. The clustering results provide insights into the common themes and patterns in students' responses. For example, the majority of transcripts within this cluster 1 are sourced from vlog 6 and vlog 8. As per the details in Table III, vlog 6 pertains to "Positionality," requiring students to elaborate on their personal stance concerning various social issues. The prompt for vlog 8 revolves around the hypothetical concept of a "Problem Solver Magic Wand," prompting students to articulate their strategies for resolving certain issues if they had access to a universal solution mechanism. The significant overlap of these vlogs within Cluster 1 suggests a possible interconnection between the subject matters of vlog 6 and vlog 8. It posits

TABLE III
VLOG KEYWORDS

Vlog	Keywords with High Probability	Priority
Vlog 1 Letter to Future Self	word, thing, time, learn, kind, work, class, hope, year, failure,	Low
Vlog 2 Motivators	feel, school, thing, motivate, kind, passion, high, confidence, people, low, experience, life, learn, work, confidence, family	Low
Vlog 3 Visualize Your Dream	people, hope, love, guess, life, work, kind, dream, feel	Low
Vlog 4 Obstacles	feel, obstacle, lot, community, work, people, government, guess, kind, dream	Medium
Vlog 5 Policy Magic Wand	access, community, homeless, build, thing, basically, policy, people, kind	High
Vlog 6 Positionality	thing, people, white, grow, community, school, people, kind, kid, lot, problem, issue	High
Vlog 7 Moment of Empathy	experience, kind, empathy, people, school, moment, friend, day	Low
Vlog 8 Problem Solver Magic Wand	kind, kid, problem, change, solution, create, issue, work, solve	High
Vlog 9 Workaround	kind, hand, chip, workaround, sister, time, implement, match, thing, work	Low
Vlog 10 Moment of Failure	moment, work, hard, year, school, time, people, failure, kind, class, thing, learn, lot	Low

that students' discourse on their position relative to social issues they care about (vlog 6) may naturally lead them to discuss potential solutions to these same issues (vlog 8). Thus, this clustering pattern could reflect the students' propensity to link their perspectives on social issues with corresponding solution-oriented discussions.

Clustering analysis also revealed interesting patterns in how individual students' responses to the vlog prompts changed over time. Table V shows the distribution of each student's transcripts across clusters. For example, Student 7's transcripts were distributed across four unique clusters. Many of the student 7's transcripts addressed the same social innovation issue, and were clustered together. Since these vlogs were submitted at different points in the semester, it suggests that student 7's transcripts talked about the same kinds of issues throughout the semester. The same is also true for student 4. In contrast, student 9 and 10's transcripts appear in eight different clusters, indicating that these students talk about many different topics over the course of the semester. The wide distribution could indicate that these students were interested in many different social innovation issues, but it could also indicate that they did not develop a sustained dialogue about one social issue. Changes in student dialogue over time are an important indicator of their evolving understandings of content in the course, as well as in shifts in their ways of participating and engaging in the class [19].

V. DISCUSSION

A major challenge of our work was the size of our dataset. Generally, in the fields of machine learning and Natural Language Processing (NLP), larger datasets tend to yield better model performance. Abundant data allows the model to recognize patterns effectively and apply them to new, unseen data. However, with a smaller data set, like ours, the model may not learn as efficiently. Due to the limited size of our dataset, we were also unable to create a supervised model, which typically requires substantial data for effective training. These limitations could be partially ameliorated in the future by collecting vlog data from larger classes, and/or to remove the time limit on students' vlog responses so that they can more fully elaborate on their ideas. Furthermore, our study adds to the growing literature that combines automated and manual methods in qualitative engineering research. A potential next step could be using Large Language Models (LLMs) like GPT. Brown et al. showed that GPT-3 can generate text that's nearly indistinguishable from human writing [20]. Even though we mainly used traditional NLP tools, using LLMs might offer deeper insights and richer context in analyzing qualitative data.

Another limitation is the potential loss of important contextual information. The pre-processing step in par-

TABLE IV
CLUSTER ANALYSIS OUTPUT

Cluster	Terms	# of vlogs
0	[‘dream’, ‘obstacle’, ‘kind’, ‘guess’, ‘thing’, ‘feel’, ‘lot’, ‘time’, ‘goal’, ‘youth’, ‘visualize’, ‘figure’, ‘stuff’, ‘work’, ‘create’, ‘number’, ‘life’, ‘future’, ‘small’, ‘live’]	10
1	[‘problem’, ‘community’, ‘issue’, ‘solve’, ‘solution’, ‘lot’, ‘people’, ‘grow’, ‘education’, ‘environmental’, ‘level’, ‘make’, ‘black’, ‘support’, ‘recognize’, ‘basic’, ‘live’, ‘reason’, ‘work’, ‘talk’]	8
2	[‘kid’, ‘foster’, ‘kind’, ‘program’, ‘lot’, ‘school’, ‘work’, ‘problem’, ‘teacher’, ‘wedding’, ‘mom’, ‘student’, ‘extremely’, ‘parent’, ‘tell’, ‘education’, ‘involve’, ‘love’, ‘difficult’, ‘dream’]	4
3	[‘day’, ‘moment’, ‘feel’, ‘friend’, ‘talk’, ‘kind’, ‘empathy’, ‘people’, ‘pandemic’, ‘understand’, ‘engineering’, ‘guess’, ‘today’, ‘school’, ‘life’, ‘work’, ‘great’, ‘drive’, ‘lose’, ‘experience’]	6
4	[‘passion’, ‘confidence’, ‘high’, ‘motivate’, ‘drive’, ‘work’, ‘low’, ‘community’, ‘school’, ‘project’, ‘accomplish’, ‘area’, ‘engineer’, ‘experience’, ‘aspect’, ‘confident’, ‘sort’, ‘social’, ‘environmental’, ‘thing’]	6
5	[‘access’, ‘policy’, ‘water’, ‘problem’, ‘community’, ‘people’, ‘basically’, ‘create’, ‘thing’, ‘opportunity’, ‘issue’, ‘role’, ‘address’, ‘technology’, ‘school’, ‘play’, ‘lot’, ‘housing’, ‘provide’, ‘represent’]	7
6	[‘human_trafficke’, ‘issue’, ‘victim’, ‘race’, ‘government’, ‘poverty’, ‘background’, ‘technology’, ‘main’, ‘people’, ‘family’, ‘guess’, ‘grow’, ‘woman’, ‘big’, ‘life’, ‘team’, ‘obstacle’, ‘person’, ‘class’]	4

TABLE V
DISTRIBUTION OF UNIQUE CLUSTERS ACROSS STUDENT VLOGS

Student	Unique number of Clusters
Student 1	7
Student 2	-
Student 3	7
Student 4	5
Student 5	5
Student 6	7
Student 7	4
Student 8	1
Student 9	8
Student 10	8

ticular is critical; for example, we had to modify the default stopword library because it initially removed words that were highly related to social innovation. Even with these corrections, it is possible that students conveyed important ideas about social innovation using terminology we were not anticipating, and lost through pre-processing. We also had to double check the topic modelling results to see how words were used in context, in order to decide whether the keywords were in fact related to students’ ideas about social innovation.

Despite these limitations, our approach, which combined topic modeling, clustering, and summarizing, did provide valuable insights. We found that NLP methods can complement manual analysis techniques by, for example, helping identify which transcripts contain the most evidence of interest and warrant more detailed qualitative analysis. These methods can also help detect patterns in transcript data over time, which is useful for longitudinal analyses of student learning within a course. Our results also add to the growing body of literature illustrating how automated methods might complement manual analysis of qualitative data engineering education research.

VI. CONCLUSION

In conclusion, the approach presented in this paper offers a unique and advantageous method for analyzing large volumes of qualitative data in engineering education research. By combining computational techniques such as topic modeling [3], cluster analysis [4], and summarization techniques, our approach provides a more efficient, objective, and scalable solution for identifying patterns and themes within transcript data. A unique aspect of our approach is the integration of topic modeling, summarizing techniques, and cluster analysis, which has not been extensively explored in prior research. This combination enables a comprehensive understanding of the data by identifying key topics, extracting important transcripts within the overall data corpus, and grouping similar transcripts together. While our approach may not fully capture the nuances and context present in qualitative data compared to manual content analysis, the

computational approach provided a fast way to prioritize the transcripts for more in-depth analysis. Overall, the approach presented in this paper demonstrates the potential for innovative combinations of computational and qualitative methods to enhance the analysis of qualitative data in engineering education research, paving the way for future research and applications in related disciplines.

REFERENCES

[1] J. W. Creswell and C. N. Poth, *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2016.

[2] L. Rourke and T. Anderson, "Validity in quantitative content analysis," *Educational technology research and development*, vol. 52, no. 1, pp. 5–18, 2004.

[3] A. K. McCallum, "Mallet: A machine learning for languagetoolkit," *http://mallet. cs. umass. edu*, 2002.

[4] D. Arthur and S. Vassilvitskii, "K-means++ the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[5] A. Nenkova, K. McKeown *et al.*, "Automatic summarization," *Foundations and Trends® in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, 2011.

[6] C. G. Berdanier, E. Baker, W. Wang, and C. McComb, "Opportunities for natural language processing in qualitative engineering education research: Two examples," in *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2018, pp. 1–6.

[7] S. Bhaduri, M. Soledad, T. Roy, H. Murzi, and T. Knott, "A semester like no other: Use of natural language processing for novice-led analysis on end-of-semester responses on students' experience of changing learning environments due to covid-19," in *2021 ASEE Virtual Annual Conference Content Access*, 2021.

[8] A. Singhal *et al.*, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[9] A. Ganesh, H. Scribner, J. Singh, K. Goodman, J. Hertzberg, and K. Kann, "Response construct tagging: Nlp-aided assessment for engineering education," in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 2022, pp. 250–261.

[10] C. A. Arbogast and D. Montfort, "Applying natural language processing techniques to an assessment of student conceptual understanding," in *2016 ASEE Annual Conference & Exposition*, 2016.

[11] F. Kolini and L. Janczewski, "Clustering and topic modelling: A new approach for analysis of national cyber security strategies," 2017.

[12] E. Shittu, D. H. B. Gai, S. LeBlanc, E. C. Wortham, and A. K. Tannon, "Uncovering strategies to improve student engagement and enhance the engineering education curriculum," in *2021 ASEE Virtual Annual Conference Content Access*, 2021.

[13] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[14] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, pp. 15 169–15 211, 2019.

[15] D. Mimno, M. Hoffman, and D. Blei, "Sparse stochastic inference for latent dirichlet allocation," *arXiv preprint arXiv:1206.6425*, 2012.

[16] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bring order to the web," technical report, Stanford University, Tech. Rep., 1998.

[18] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[19] M. Ryu and T.-R. Sikorski, "Tracking a learner's verbal participation in science over time: Analysis of talk features within a social context," *Science Education*, vol. 103, no. 3, pp. 561–589, 2019.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.