Natural Language Processing for Theoretical Framework Selection in Engineering Education Research

Catherine G.P. Berdanier
Department of Mechanical Engineering
The Pennsylvania State University
University Park, PA 16802
cgb9@psu.edu

Christopher M. McComb
School of Engineering Design and Professional Practice
The Pennsylvania State University
University Park PA, 16802
uum209@psu.edu

Weiwei Zhu
Department of Mechanical Engineering
The Pennsylvania State University
University Park, PA 16802
wzz46@psu.edu

***Abstract*— This research paper presents recent work exploring the power of natural language processing (NLP) methods applied to qualitative engineering education data. As NLP and other machine learning methods are developed for qualitative data, it is important to prioritize the role that theory plays in rigorous qualitative research, where the selection of a theoretical framework serves as the lens by which the research project is framed, results are analyzed, and findings are brought to light. Indeed, the view from a different theoretical lens can highlight novel or new findings. In this work, we seek to explore the viability of NLP methods for helping researchers select appropriate frameworks. In this work, we present our method to train a Python-based NLP algorithm to analyze an existing data set of interview data using one theoretical lens: Community of Practice theory, an oft-used theory in graduate education literature, which is the topic of the interview corpus to investigate. We present and test two methods for developing dictionaries by which to train the algorithm: An expert-curated dictionary and a machine-generated dictionary compiled by mining the theoretical framework sections of published literature employing Community of Practice theory. We apply these two dictionaries to analyze a corpus of 54 interview transcripts investigating graduate engineering attrition. The high dimensional data from NLP can be compared using Principal Component Analysis (PCA) visualization and pairwise distance plots to determine which method results in the most well-defined structure indicating agreement between the dictionary and the corpus of interview transcripts. In the discussion, we highlight opportunities for using these automated methods to help researchers with qualitative data analysis and warn against potential dangers and ethical ramifications for using machine learning and NLP for social science data. This work will have impact on the disciplinary communities working to embed computational language-based methods into engineering education research, and for the qualitative methods communities across social science and education disciplines.**

Keywords—Natural Language Processing, Engineering Education, Theoretical Frameworks

I. Introduction

The inclusion of theoretical frameworks in social science and educational research is important to provide a lens through which data—whether quantitative or qualitative, can be interpreted [1], [2]. The selection of a theoretical framework can illuminate certain aspects of the data, while veiling others—while the selection of a different framework might illuminate other facets of a given phenomenon. There are an abundance of theories related to education, psychology, sociology, and other fields that likely would yield useful or novel interpretations on engineering education findings, but many researchers rely on the theoretic orientations with which they are most comfortable, causing the field to go through trends in the use of various theories. The motivation for this paper is grounded in a visionary perspective on the ability to use theory in engineering education—and other disciplines employing human subjects research, too—leveraging the power of machine learning to help researchers select interesting, novel, and useful theoretical frameworks to illuminate the answers to research questions. We envision—in the long term—a tool to help researchers compare their corpus of qualitative text-based data (interviews, etc.) against a repository of theories from across disciplines to help them understand their data in new or novel ways.

The purpose of this paper is to present a proof-of-concept using natural language processing tools to develop dictionaries representing one theory (Lave and Wenger's Community of Practice Theory [3], commonly applied in engineering education research). We use this research to pilot and compare two ways of formulating a working dictionary by which to apply natural language processing methods: an expert-curated dictionary and a machine-generated dictionary. Then, we show how the dictionaries can be applied to a corpus of interview data to assess the relevancy of the theory to the corpus.

II. Literature Review

Natural language processing (NLP) is a method complimentary to machine learning, in which researchers ask computers to determine underlying hidden structure or patterns in large text-based data sets. NLP, in particular, is employed in

a variety of daily technologies that anticipate language patterns based on rhetorical structure, such as search results on the internet or voice recognition in artificial intelligence. In educational settings, NLP has been a topic of interest with the advent of online learning and Massive Open Online Courses (MOOCs), with research groups working in the area of automated grading and feedback [4]–[7] or in analysis of student written essays [8].

In engineering education research, few researchers employ machine learning or NLP. Madhavan and Johri's collaborations [9]–[11] have employed machine learning, big data visualization methods, and content analysis to discern trends in engineering education literature and research topics. Multiple groups employed NLP techniques to analyze large batches of student data. For example, Meneske's group used NLP methods to explore relationships between student goals, reflections, and learning outcomes[12], [13], and Bhaduri [14] showed how NLP methods could be used to interpret student metacognition and course evaluations, with the goal of helping instructors synthesize large amounts of text-based data. Several researchers have applied NLP to the teaching of engineering skillsets, where NLP-based algorithms can act as tutors for computational courses [15] or for writing [16]. In the disciplinary design education domain (not engineering design thinking), several researchers have been employing machine learning and AI to the design process as a way of better characterizing design in a variety of context or generating robust design repositories digitally. In our past work, we have used NLP for analysis of engineering résumés [17] and other machine learning methods to study engineering design characteristics in a variety of settings [18].

Despite these ongoing projects, however, the use of NLP or ML is not widespread engineering education research. One potential reason for this is that by asking a machine to help with data analysis, some may worry that the richness of qualitative methods that come from theory-based inductive reasoning, especially for deep qualitative methods, may lose the creativity, abductive and inductive reasoning that a human researcher provides, and may not offer rich advances on psychological or sociological human conditions. We fully agree, but also note that researchers will likely be applying NLP and ML methods to qualitative data whether researchers concur it to be sound or not. Indeed, in the last few years there have been several papers published employing NLP for qualitative data analysis. We note a few of the most useful examples here.

Crowston, Allen, and Heckman [19] were some of the first researchers to apply NLP methods to qualitative data analysis, focusing on methods of content analysis, which is focused on frequency counting and thematic analysis. They demonstrated the ability of leveraging NLP to extract codes from text, maximizing the amount of time required by human coders. Of note is that the purpose of this paper is to "explore how NLP techniques […] can be applied to support a particular task in positivist qualitative research, namely coding for content analysis" (p. 524). The obvious limitation here is that most human-centric and educational research has shifted to being post-positivist and constructivist in nature, and that more nuanced interpretive qualitative methods are not addressed.

Guetterman et al. [20] quantified the difference between an NLP-enabled coding schema with human qualitative analysis, finding that, unsurprisingly, the NLP-trained algorithm yielded three main themes of analysis (compared with four from the human analysis), while missing aspects of context. The team also ran an experiment with an "augmented" (human + NLP) coding schema, noting that the combination of humans with machine-assisted techniques could allow researchers to handle more data while not losing the contextual features of qualitative data. However, this group did not incorporate or address aspects of theoretical orientation within their paper.

Chen et al. [21] also discussed the use of NLP and ML to support qualitative data analyses in social science research, identifying a strength in using ML and related methods to *identify* areas of ambiguity that then could be followed up with targeted researcher analysis and interpretation. This team also points to a future of human-centered machine learning, rather than a world where a machine can be fully able to replace humans in qualitative research. Chen et al. also note that one of the main difference between most ML-based applications and that of social science is that it is driven by theory, with the exception of grounded theory methods [22], which originate with the presumed absence of theory, noting that ML could contribute to these grounded theory methodologies.

In each of these papers, the essential aspects of theory to social science are ignored, which is a crucial aspect if ML or NLP methods will ever be accepted in social science and educational communities. In the discipline of engineering education research, the importance of theory is well-established, and required in our premiere research journals. To this end, our team seeks to be on the front end of the wave to propose methods for applying NLP to engineering education and other qualitative-heavy disciplines that support underlying traditions of qualitative research. One of the most important pieces of this is our commitment to the value of theory as it serves as the lens for interesting data analysis and interpretation, leading to advances in understanding of the phenomenon of interest.

To support this goal, the purpose of this paper is to show how theoretical orientations can be trained into NLP algorithms to help researchers analyze qualitative text-based data, such as that from interviews. While we demonstrate this for one theory in this paper, the proof-of-concept can be easily expanded to help researchers quickly explore how well a variety of diverse theories might fit a corpus of data, helping researchers to expand outside their "favorite" theories. The specific overarching research question this paper seeks to address is:

How can NLP methods be used to incorporate theoretical frameworks into the analysis of text-based qualitative data sets?

III. Methods

The methods for this paper are summarized overall before giving the specifics. First, two dictionaries were generated to represent Community of Practice Theory: an expert-curated dictionary of words and phrases affiliated with Community of Practice Theory, and a machine-generated dictionary based on theory sections from 14 educational journal articles primarily employing that theory. Second, we apply these dictionaries through NLP techniques to a corpus of data, the topic of which is graduate engineering student attrition, persistence, and career trajectories. Third, we compare the results using PCA visualization and pairwise-distance plots, which represent the structure present in high-dimensional data.

A. Introduction to Context for Corpus of Data and Theory of Interest

The context for the existing corpus of interview (text-based) data comes from two funded NSF grants that both investigate different dimensions of graduate engineering student attrition and persistence from the PhD. While the studies had different foci and research questions, the semi-structured interview protocols were quite similar. In sum, the corpus comprises 54 interviews with current and former engineering PhD students at various stages of their programs. More details on recruitment of participants are available in our other published works [23], [24]; the details of which are less relevant here because the focus of this paper is the proof of concept for theoretical framework selection using NLP.

In this paper, we have selected Community of Practice Theory, originally developed by Lave and Wenger [3] to describe the development of expertise as one progresses from being peripherally associated as a member of a discipline to being a core member of the community of practice. One of the key phrases in Community of Practice theory is the concept of "legitimate peripheral participation," or the need for novices to take on gradually more authentic learning opportunities such that they develop a sense of belonging and identity as a member of the discipline. CoP is often associated with graduate level and higher education because of its focus on learning in authentic situations outside the classroom and its emphasis on apprenticeship to learn the knowledge and skills of the trade, and to internalize the expectations and norms of the discipline [25]–[30]. In engineering education research, CoP is one of the most widely-applied theories affiliated with graduate education work, and was therefore selected for this proof-of-concept study.

B. Development of Dictionaries and Cleaning of the Corpus of Interview Data

In this section introduce the methods by which the dictionaries were generated and the NLP models were created. The first dictionary is an "expert-curated dictionary" comprising 69 words and phrases, generated by the authors, who have extensive experience in graduate education research employing Community of Practice Theory. These words were saved as a .txt file to be readable into Python.

The second dictionary is a machine-generated dictionary that was formed by mining the theoretical framework sections of fourteen journal articles in both engineering education research and other higher education venues that have a thorough discussion of theoretical orientation. Depending on the venue, this sometimes exists as a stand-alone section; in others, it is embedded as a subsection or several well-developed paragraphs within the literature review or background section. In generating the set of literature from which the theory sections were mined, we excluded literature related to online graduate students and online communities of practice, since the context for the corpus against which our dictionaries would be tested was based on traditional resident doctoral programs, not online degrees. The relevant sections or paragraphs from each article were pasted into .txt files to be read into Python.

After developing the dictionaries, we also prepared our corpus of interview data to be analyzed in Python. For this process, all interviews were saved as .txt files and cleaned to remove conversations that happened either before (such as consent etc.) or after (niceties, requests for follow-up participation, logistics about how the participation incentive would be sent, etc.) were deleted from the interview transcripts, such that the algorithm would not be searching these parts of the conversation for evidence of CoP theory. We did not remove interviewer questions to the participant, or timestamps: the timestamps would not have high enough occurrences to be important to the algorithm (or we could program the algorithm to remove numeric timestamp values). We made the methodological decision to retain interviewer questions because of the semi-structured nature of the interview because the interviewer questions would help build in elements of context to the results as the algorithm.

C. Application of Dictionaries to Corpus of Data

The dictionaries were applied to analyze the corpus of data through Latent Semantic Analysis (LSA) [31]. LSA provides a means of converting text-based documents into numerical vectors, which then permits a variety of algebraic comparisons and manipulations to support various experimental needs [32]–[34].

In essence, the first step involves expressing each document in accordance with a dictionary. Here, two dictionaries are used – a data-driven dictionary produced by mining the theoretical framework sections of other work, and an expert-curated dictionary generated by the authors. Specifically, every document is represented as a frequency count over the words in the dictionary. This frequency count is then represented as a vector. Every vector are equal in length to the number of words in the dictionary, and every value in a given vector indicates how many times the corresponding dictionary word appears in the document. Singular value decomposition is then used to compress these high-dimensional vectors into a lower-dimensional latent space. In the absence of a parametric study, the number of dimensions in the latent space is made equal to the number of documents in the corpus (here, 53 document and dimensions). This step has the benefit of removing some of the noise from the empirical data. The vector representations of documents within this LSA latent space can then be used for algebraic manipulation and comparison. The nature of the latent

space is such that documents that are nearby one another are more similar in content, while those that are further away are dissimilar. In this work, two separate latent spaces are constructed – one corresponding to the data-driven dictionary, and the other corresponding to the expert-curated dictionary.

Specifically, the analysis implemented in this work was implemented using the Gensim Python library [35]. This application of the combination of LSA and visualization approaches to assess textual data is similar to that employed in other work [36].

IV. Results

This section provides visualizations of the LSA treatment of the data described above. These visualizations are of two types. First, PCA was applied to the data in order to find two principal components which best describe the variance in the data. This variance-preserving visualization approach provides an effective means of visualizing high-dimensional data (in this case, 53-dimensional) while preserving information. Second, pairwise-distance plots were produced. These plots show the distance between documents in the analysis, visualized as a matrix.

The PCA visualizations of the analysis predicated on the data-driven and expert-curated dictionaries are provided in Figure 3 and Figure 4, respectively. In these plots, every point represents a separate document. The axes and the units on the axes are not meaningful, as they are simply the components that preserve maximum variance in the data. In general, these visualizations reveal similar structures in the data. Namely, both plots have a tight grouping of documents centered at the origin, indicating a set of documents that are minimally-differentiated by the analysis. The grouping is slightly less dense for the expert-curated dictionary, which may be indicative of a greater degree of nuance enabled by that dictionary. There are also several outlying documents, all situated on the right side of the plot. This indicates documents that may have unique qualitative attributes when viewed through the lens of Community of Practice theory. Although not explored fully here, these outliers could be used to select documents for deeper qualitative review.

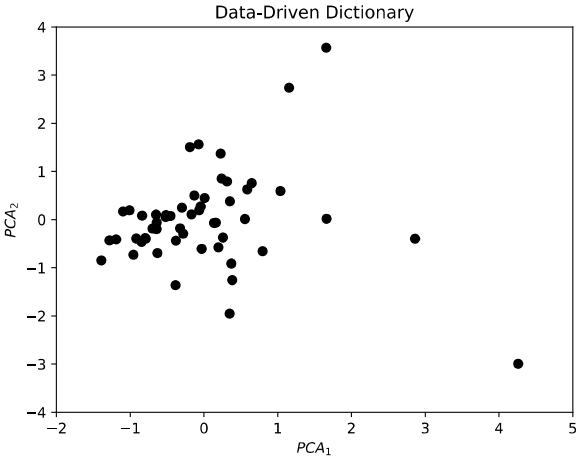

Figure 1. PCA visualization of the analysis predicated on the data-driven dictionary.

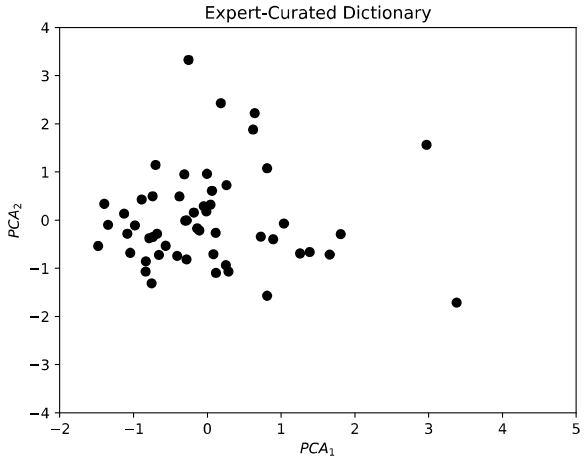

Figure 2. PCA visualization of the analysis predicated on the expert-curated dictionary.

The pairwise distance plots of the analysis predicated on the data-driven and expert-curated dictionary are provided in Figure 3 and Figure 4, respectively. These plots show the matrix of distances (as measured using Euclidean distance in the 53-dimensional LSA latent space) between each pair of documents. The matrix has zeroes along the diagonal, indicating the comparison of a document to itself. Further, both matrices are symmetric across the diagonal, simply indicative of the symmetry requirement for all distance metrics. The exact value of the distances is not meaningful, as they are produced for two separate LSA latent spaces. However, the patterns of high and low values shown in the matrices are meaningful, and further reveal the structure of the two spaces.

A heuristic comparison of the two plots indicates a high degree of similarity. This is particularly notable when observing the degree to which documents 39 and 48 are outliers, as indicated by the high values in those rows and columns. However, the more specific nature of these matrices differs. Specifically, the pairwise distance plot produced for the data-driven dictionary is composed of many very low values and a few high values. This indicates many documents that are very close to one another in the LSA latent space, with a few documents that are outliers. In contrast, the pairwise distance plot produced for the expert-curated dictionary has less of a dichotomous split, with more mid-range values. This belies a diffuse spread of documents in the LSA latent space, and is indicative of a greater degree of nuance induced by the expert-curated dictionary.

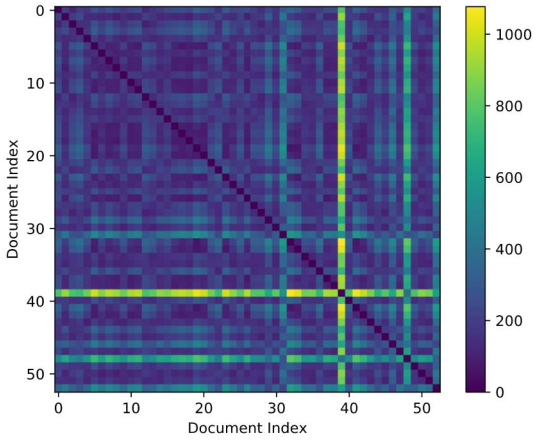

Figure 3. Pairwise-distance plot of the analysis predicated on the data-driven dictionary.

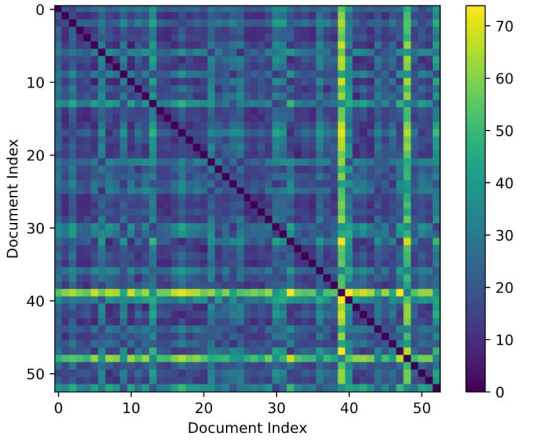

Figure 4. Pairwise-distance plot of the analysis predicated on the expert-curated dictionary.

V. Discussion and Future Work

In this work, we demonstrated the ability to train and apply an NLP algorithm to a corpus of text-based interview data. This is, to the best of our knowledge, the first documented instance of the incorporation of theoretical frameworks to qualitative analysis using NLP, offering a substantial advance to the disciplines of engineering education and to methodological communities. In this work, we applied two different dictionaries to the algorithm, one generated by experts in applying CoP theory to graduate education, and one generated by a machine based on the theoretical framework sections of peer-reviewed journal articles.

As shown by the structure present in the PCA visualizations and pairwise distance plots. While both dictionaries produced data with similar structure, the expert-curated dictionary produced fewer extreme outliers and thus provided more nuance in the treatment of the documents under analysis. While future work would need to validate this pattern for several other theories, we now have a starting point to be able to judge whether an expert-curated dictionary or a machine-generated dictionary is most appropriate to analyze a data set.

This research was a first step in developing theory-driven NLP algorithms to help qualitative researchers analyze data. Because of this, we selected the theoretical framework of CoP that is often employed to describe graduate socialization through the apprenticeship, thereby giving us an equal ground by which to compare how the two dictionaries represented the data.

There are several theoretical ramifications that researchers should hold in their minds as they consider using NLP or other ML methods for qualitative data analysis. First, we hold strongly to the idea that algorithm-based methods of analysis should never take the place of human researchers, especially in qualitative analysis where the interpretation of "messy" human data is often nuanced and not well-described by an existing theory. Indeed, in many contexts, a combination of theories are used to describe a phenomenon, and in exceptional circumstances, grounded theory methods are employed when there is no theory that can describe the context explored in a research study. It is important for researchers to have a deep appreciation and understanding for qualitative traditions and applications of theory to data before attempting to build NLP or ML algorithms to begin to help analyze or interpret data. Without the researcher's understanding and expertise, then that leaves a machine to tell the research community the findings of any given research project: an inappropriate use of technology.

We also warn against the tendency to consider machine-generated data as "objective." Indeed, machines and algorithms are only as smart as the programmers, and can serve to propagate bias under the pretense of objectivity. Recent reports have shown how bias propagates in online knowledge repositories such as Wikipedia (for example, which covers male scientists, inventors, and technologists, at higher levels than females). Problematically, many search engines are trained on Wikipedia articles, such that the representation bias manifests in considering science, engineering, and technology to be affiliated with males rather than females. This is just one example of how bias pervades in an invisible circumstance, while no single "person" is responsible for that bias. We must be careful to keep the researcher in the process, and to acknowledge the limitations of these methods and to confront positionalities of the researchers if NLP or machine learning methods are employed to conclude findings from qualitative research.

Future work for this project includes further honing how dictionaries should be built. We intend to compare the two dictionaries built for this study, for example, with "augmented" versions of the dictionaries enhanced by coding Python to expand the dictionary to include synonyms for all the words and phrases within each of our existing dictionaries. Further, we intend to build a repository of other theories that are commonly (and less-commonly) used in engineering education work, and apply them to a variety of data sets to show how these methods might work to help researchers decide on a most appropriate theoretical framework to help guide their work.

VI. Conclusions

In this paper, we demonstrated the incorporation of theoretical orientations within natural language processing, demonstrating that for the theory of interest for this paper (Community of Practice Theory applied to a corpus of interview data pertaining to graduate-level socialization and attrition) the dictionary developed by the expert was a better fit. This is the first step to developing methods and best practices for appropriately using NLP in qualitative data analysis methods, keeping in mind the ethical and methodological traditions that should continue to be highly valued in qualitative research. In the discussion, we offer suggestions and cautions to qualitative researchers as more and more disciplines are beginning to capitalize upon the advantages of ML and NLP.

Acknowledgment

We thank the teams of researchers who collected the interview data that was used for this study, specifically Ellen Zerbe, Kanembe Shanachilubwa, and Gabriella Sallai. This material is based upon work supported by the National Science Foundation under Grants 1844878 and 1733594. 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] A. Youngblood Jackson and L. Mazzei, Thinking with theory in qualitative research: Viewing data across multiple perspectives. Routledge, 2011.

[2] G. J. Mitchell and W. K. Cody, "The role of theory in qualitative research," *Nurs. Sci. Q.*, vol. 6, no. 4, pp. 170–178, 1992.

[3] J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*. Cambidge University Press, 1991.

[4] A. Shehab, M. Elhoseny, and A. E. Hassanien, "A hybrid scheme for automated essay grading based on LVQ and NLP techniques," *2016 12th Int. Comput. Eng. Conf. ICENCO 2016 Boundless Smart Soc.*, pp. 65–70, 2017.

[5] V. M. Holland and J. D. Kaplan, "Natural language processing techniques in computer-assisted language learning: Status and instructional issues," *Instr. Sci.*, vol. 23, no. 5–6, pp. 351–380, 1995.

[6] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, "Forecasting student achievement in MOOCs with natural language processing," *ACM Int. Conf. Proceeding Ser.*, vol. 25-29-April-2016, pp. 383–387, 2016.

[7] S. Crossley, D. S. Mcnamara, L. Paquette, R. S. Baker, and M. Dascalu, "Combining click-stream data with NLP tools to better understand MOOC completion," *ACM Int. Conf. Proceeding Ser.*, vol. 25-29-April-2016, pp. 6–14, 2016.

[8] S. Crossley, D. Russell, K. Kyle, and U. Römer, "Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields?," *J. Writ. Anal.*, vol. 1, no. 1, pp. 48–81, 2017.

[9] A. Johri, G. A. Wang, X. Liu, and K. Madhavan, "Utilizing topic modeling techniques to identify the emergence and growth of research topics in engineering education," *Proc. - Front. Educ. Conf. FIE*, pp. 1–6, 2011.

[10] K. Madhavan, H. Xian, A. Johri, M. Vorvoreanu, B. K. Jesiek, and P. C. Wankat, "Understanding the Engineering Education Research problem space using Interactive Knowledge Networks," *ASEE Annu. Conf. Expo. Conf. Proc.*, 2011.

[11] K. Madhavan, A. Johri, H. Xian, G. A. Wang, and X. Liu, "Tools for large-scale data analytic examination of relational and epistemic networks in engineering education," *Adv. Eng. Educ.*, vol. 4, no. 2, pp. 1–36, 2014.

[12] D. Heo, S. Anwar, and M. Menekse, "How do engineering students' achievement goals relate to their reflection behaviors and learning outcomes?," *ASEE Annu. Conf. Expo. Conf. Proc.*, vol. 2017-June, 2017.

[13] D. Heo, S. Anwar, and M. Menekse, "The relationship between engineering students' Achievement goals, reflection behaviors, and learning outcomes," *Int. J. Eng. Educ.*, vol. 34, no. 5, pp. 1634–1643, 2018.

[14] S. Badhuri, "NLP in engineering education: Demonstrating the use of Natural Language Processing techniqes for use in Engineering Education classrooms and research," Dissertation: Virginia Tech, 2017.

[15] S. Müller, B. Bergande, and P. Brune, "Robot tutoring: On the feasibility of using cognitive systems as tutors in introductory programming education: A teaching experiment," *ACM Int. Conf. Proceeding Ser.*, pp. 45–49, 2018.

[16] S. Ghosh, "Online automated essay grading system as a web based learning (WBL) tool in engineering education," *Web-Based Eng. Educ. Crit. Des. Eff. Tools*, pp. 53–62, 2010.

[17] C. G. P. Berdanier, E. Baker, W. Wang, and C. McComb, "Opportunities for Natural Language Processing in Qualitative Engineering Education Research: Two Examples," in *IEEE Frontiers in Education*, 2018.

[18] P. U. Mehta, M. Malviya, C. M. McComb, G. P. Manogharan, and C. G. P. Berdanier, "Mining design heuristics for additive manufacturing via eyetracking methods and hidden markov modellin." *Submitted to ASME J. Mech Design.*

[19] K. Crowston, E. E. Allen, and R. Heckman, "Using natural language processing technology for qualitative data analysis," *Int. J. Soc. Res. Methodol.*, vol. 15, no. 6, pp. 523–543, 2012.

[20] T. C. Guetterman, T. Chang, M. DeJonckheere, T. Basu, E. Scruggs, and V. G. V. Vydiswaran, "Augmenting qualitative text analysis with Natural Language Processing: Methodological study," *J. Med. Internet Res.*, vol. 20, no. 6, p. e231, 2018.

[21] N. C. Chen, M. Drouhard, R. Kocielnik, J. Suh, and C. R. Aragon, "Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 2, 2018.

[22] K. Charmaz, *Constructing grounded theory: A practical guide through qualitative research.* London: SAGE Publications, Inc., 2006.

[23] E. Hocker, E. Zerbe, and C. G. P. Berdanier, "Characterizing Doctoral Engineering Student Socialization: Narratives of Mental Health, Decisions to Persist, and Consideration of Career Trajectories," in *IEEE Frontiers in Education*, 2019, pp. 1–7.

[24] E. Zerbe and C. G. P. Berdanier, "Writing Attitudes and Career Trajectories of Domestic and International Students in the United States," *Int. J. Eng. Educ.*, vol. 36, no. 1A, pp. 226–240, 2020.

[25] V. L. O'Donnell and J. Tobbell, "The transition of adult students to higher education: Legitimate peripheral participation in a community of practice?," *Adult Educ. Q.*, vol. 57, no. 4, pp. 312–328, 2007.

[26] L. K. Newswander and M. Borrego, "Using journal clubs to cultivate a community of practice at the graduate level," *Eur. J. Eng. Educ.*, vol. 34, no. 6, pp. 561–571, 2009.

[27] E. Crede, M. Borrego, and L. D. McNair, "Application of community of practice theory to the preparation of engineering graduate students for faculty careers," *Adv. Eng. Educ.*, vol. 2, no. 2, pp. 1–22, 2010.

[28] H. Y. Kim, "International graduate students' difficulties: Graduate classes as a community of practices," *Teach. High. Educ.*, vol. 16, no. 3, pp. 281–292, 2011.

[29] M. Nerad, "Conceptual approaches to doctoral education: A community of practice," *Alternation*, vol. 19, no. 2, pp. 57–72, 2012.

[30] K. Coffman, P. Putman, A. Adkisson, B. Kriner, and C. Monaghan, "Waiting for the expert to arrive: Using a community of practice to develop the scholarly identity of doctoral students.," *Int. J. Teach. Learn. High. Educ.*, vol. 28, no. 1, pp. 30–37, 2016.

[31] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, 1990.

[32] K. Fu, J. Cagan, and K. Kotovsky, "Design team convergence: The influence of example solution quality," *J. Mech. Des.*, vol. 132, no. 11, p. 111005, 2010.

[33] A. Dong, A. W. Hill, and A. M. Agogino, "A document analysis method for characterizing design team performance," *J. Mech. Des.*, vol. 126, no. 3, pp. 378–385, 2004.

[34] A. Dong, M. S. Kleinsmann, and F. Deken, "Investigating design cognition in the construction and enactment of team mental models," *Des. Stud.*, vol. 34, no. 1, pp. 1–33, 2013.

[35] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," *Proc. Lr. 2010 Work. New Challenges NLP Fram.*, 2010.

[36] C. McComb and F. M. Tehrani, "Research and practice group methodology: A case study in student success," in *2014 ASEE Pacific-Southwest Conference*, 2014.

.