NLP IN ENGINEERING EDUCATION

Demonstrating the use of Natural Language Processing techniques for use in Engineering Education classrooms and research

Sreyoshi Bhaduri

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Engineering Education

Holly M. Matusovich, Chair

David B. Knight

Elizabeth D. McNair

Glenda R. Scales

December 20, 2017

Blacksburg, VA

Keywords: Machine learning, Natural Language Processing, Engineering Education, Exploratory Qualitative Research

NLP IN ENGINEERING EDUCATION

Demonstrating the use of Natural Language Processing in Engineering Education classrooms and research

Sreyoshi Bhaduri

ABSTRACT

Engineering Education is a developing field, with new research and ideas constantly emerging and contributing to the ever-evolving nature of this discipline. Textual data (such as publications, open-ended questions on student assignments, and interview transcripts) form an important means of dialogue between the various stakeholders of the engineering community. Analysis of textual data demands consumption of a lot of time and resources. As a result, researchers end up spending a lot of time and effort in analyzing such text repositories. While there is a lot to be gained through in-depth research analysis of text data, some educators or administrators could benefit from an automated system which could reveal trends and present broader overviews for given datasets in more time and resource efficient ways. Analyzing datasets using Natural Language Processing is one solution to this problem.

The purpose of my doctoral research was two pronged: first, to describe the current state of use of Natural Language Processing as it applies to the broader field of Education, and second, to demonstrate the use of Natural Language Processing techniques for two Engineering Education specific contexts of instruction and research respectively. Specifically, my research includes three manuscripts: (1) systematic review of existing publications on the use of Natural Language Processing in education research, (2) automated classification system for open-ended student responses to gauge

metacognition levels in engineering classrooms, and (3) using insights from Natural Language Processing techniques to facilitate exploratory analysis of a large interview dataset led by a novice researcher.

A common theme across the three tasks was to explore the use of Natural Language Processing techniques to enable the computer to extract meaningful information from textual data for Engineering Education related contexts. Results from my first manuscript suggested that researchers in the broader fields of Education used Natural Language Processing for a wide range of tasks, primarily serving to automate instruction in terms of creating content for examinations, automated grading or intelligent tutoring purposes. In manuscripts two and three I implemented some of the Natural Language Processing techniques such as Part-of-Speech tagging and tf-idf (text frequency- inverse document frequency) that were found (through my systematic review) to be used by researchers, to (a) develop an automated classification system for student responses to gauge their metacognitive levels and (b) conduct an exploratory novice led analysis of excerpts from interviews of students on career preparedness, respectively. Overall results of my research studies indicate that although use of Natural Language Processing techniques in Engineering Education is not widespread, although such research endeavors could facilitate research and practice in our field. Particularly, this type of approach to textual data could be of use to practitioners in large engineering classrooms who are unable to devote large amounts of time to data analysis, but would benefit from algorithmic systems that could quickly present a summary based on information processed from available text data.

NLP IN ENGINEERING EDUCATION

Demonstrating the use of Natural Language Processing in Engineering Education classrooms and research

Sreyoshi Bhaduri

GENERAL AUDIENCE ABSTRACT

Textual data (such as publications, open-ended questions on student assignments, and interview transcripts) form an important means of dialogue between the various stakeholders of the engineering community. However, analyzing these datasets can be time consuming as well as resource-intensive. Natural Language Processing techniques exploit the machine's ability to process and handle data in time-efficient ways. In my doctoral research I demonstrate how Natural Language Processing techniques can be used in the classrooms and in education research. Specifically, I began my research by systematically reviewing current studies describing the use of Natural Language Processing for education related contexts. I then used this understanding to inform use of Natural Language Processing techniques to two Engineering Education specific contexts: one in the classroom to automatically classify students' responses to open-ended questions to understand the metacognitive levels, and the second context of informing analysis of a large dataset comprising excerpts from interview transcripts of engineering students describing career preparedness.

My research shows that although use of Natural Language Processing techniques in Engineering Education is not widespread, use of such techniques could facilitate research and practice in this field. Particularly, this type of approach to textual data could be of use to practitioners in engineering classrooms who are unable to devote large amounts of time to data analysis, but would benefit from a system that is able to provide them feedback about their instruction through analyzing student reflective responses.

Dedicated to my parents:
Brigadier (Dr.) Soumyesh Nath Bhaduri & Dr. Indrani Bhaduri

ACKNOWLEDGEMENTS

When I was five years old, both my parents earned their doctoral degrees. I recall hours of furious typing followed by the buzzing as the dot matrix printer printed out bundles of chapters of their dissertations, only to go through yet another round of red-inked edits and rework. Months of this, as they worked full time, and tended to a menagerie with (at that time) twelve cats, two pups, a cow, and me. At that time, I vicariously experienced the dissertation process, twice, both of which had seemed to be all fun and large bundles of printouts. This time around, my parents re-lived their dissertation through mine. Thank you, Baba and Maa, for being a part of my doctoral journey, just as I was yours. Thank you also for being a constant source of inspiration, and for believing in me, and loving me: always, anyways. This one, like most of what I do, would not have been possible without you (and Simba).

I am also grateful for every individual who supported, encouraged, and facilitated my growth over the past couple of months. Most importantly, Dr. Holly M. Matusovich: my advisor, mentor, and sounding board; for all those hours she took to painstakingly help me mold my abstract ideas into concrete manuscripts. Thank you for guiding me through this dissertation process, for always being my advocate, for teaching me to be more considerate, mindful, and organized, and for being a role model of an advisor and academician. This document is a testament to your infinite patience and belief in me.

Dr. Glenda Scales: I recall my first meeting with you, where you urged me to always remember why I was passionate about this idea to begin with, and to “enjoy the doctoral journey, ‘coz it only comes once”. Thank you for your never-ending support, your warmth, and your unique perspectives on not only my dissertation but also research in general. Dr. Lisa Mc. Nair: for being such an inspiration when it comes to thinking out of the box, being interdisciplinary, and doing it all with a smile. Thank you for your deep questions, and for the discussions based on your linguistically inclined interpretations of my results which helped me learn to think more deeply about words I use and their

meanings. Dr. David Knight: for being honest, pragmatic, and incredibly detailed in your feedback. I am incredibly lucky to have had the opportunity to work with you on projects beyond my dissertation. Thank you not only for the guidance, but also for leading by example. Although not on my committee, I also learned from my interactions and work with Dr. Walter Lee, Dr. Anne Ryan Driscoll, Dr. Elizabeth Creamer, Dr. Michael von Spakovsky and Dr. Richard Goff: each of whom contributed to my growth as a mixed methods researcher, a data enthusiast, an instructor, an engineer, and overall, an individual, in their own unique ways.

The dissertation process with all its documentations and deadlines would definitely not have been smooth sailing without Linda Hazelwood. The same holds true for all the staff at the Graduate School. Thank you for everything you do!

I would probably not have pursued a Ph.D. in Engineering Education if I had not learned about the differences and similarities in higher education and associated challenges across the globe, through my interactions with Dean Karen DePauw and the GPP 2013 cohort. Thank you Dean DePauw, for teaching me the importance of creating safe and brave spaces wherever I go, and for understanding and valuing diversity and striving for inclusivity. Thank you also to the wonderful graduate school community, especially my colleagues and friends through VTGrATE, GSA, and GPP: especially Nicole Johnson, Tara Reel and Shaima Abdallah. My doctoral journey was a continuation of my Graduate School life in the town of Blacksburg. I am immensely thankful to this whimsical town and the wonderful community for all the love it has showered on me through these years, especially Krisha Chachra and April Amodeo for their love, warmth and friendship.

I am also thankful to the family I found at Blacksburg: especially perennial love from Riddhika Jain, and Gaurav Soni, life-hacks from Richa Sinha, music, conversation and shared laughter with Shubham Chowdhury (and Manni), Pratik Anand, and Shuvodip Bhattacharya, and frequent doses of fraternal support from Sarankumar Venkatapathi,

and Satyajit Upasani. Thank you for making these six years away from home, ‘almost’ home. In addition, I will eternally be indebted to Hrusheekesh Warpe for re-igniting in me my passion for literature, Austen, poetry and poets, and for helping me find magic in the mundane: “*Medicine, law, business, engineering, these are noble pursuits and necessary to sustain life. But poetry, beauty, romance, love, these are what we stay alive for*.”

I am blessed also to have had perpetual love from Zoya Dixit and Saloni Sharma, my life-long friends, who continue to inspire me by being confident, hard-working women with a zeal for life and adventure. I am also thankful for a new friend, Sanjukta Sarkar, for her cheery wit, infinite wisdom and her fierce, inspiring, independence, which I hope to continue to learn from. I am also thankful for Dr. Lily Virguez (and Dr. Bambam!), Cynthia Hampton, Dr. Cherie Edwards, Desen Ozken, Michelle Soledad, Adetoun Taiwo, and Debarati Basu: incredibly brilliant, kind and wonderful women in the Department of Engineering Education, for sharing parts of their doctoral journey with me, and for contributing to bettering the world one dissertation at a time.

It was extremely helpful to have had a wonderful cohort within the Department of Engineering Education, and insightful mentors in Dr. Homero Murzi, Dr. Glenda Young, and Dr. Brian Novoselich. Additionally, this thesis benefited from help and all the resources of the University Library: especially, Philip Young and Nathan Hall. A shout out also to members of SMILE research group, past and present. Especially, Dr. Cheryl Carrico: thank you for being fabulous, and for the never-ending positivity you brought to my doctoral process, at every step.

Dr. Tamoghna Roy: We did it! Thank you for sharing your doctoral journey with me, and for being the Tommy to my Tuppence. Thank you also for the numerous drives, for bearing with my dramatic renditions of every situation and the probable outcomes, for the late-night writing/editing marathons (often in higher proportion to our Psych binges), for

all the packages of books, bouquets of sharpened pencils, and cheery post-its, but above all for being my closest friend and confidante through this entire process.

It was difficult to shorten the paragraphs as I tried to remember to thank all the wonderful persons who were important to me through this rather eventful journey. However, I am sure to have missed out many names. So, thank you, to everyone who helped me learn and grow.

Finally, thank you to Agatha Christie, Satyajit Ray and Sylvia Plath, for their work, because my doctoral journey often necessitated rewinding with hot chocolate and some (unrelated, non-academic, comfort) reading.

“Time flies like an arrow, fruit flies like a banana”

- Groucho Marx

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 OVERVIEW

Growing class sizes and increased enrollment in universities have led to increases in the volume of data being generated in education (Parry, 2012; Soledad et al., 2017). This increase in volume has in turn led to several challenges for education researchers and instructors. For example, providing timely and effective feedback to students in large classrooms can be difficult (Soledad & Grohs, 2017). My research addresses some of the challenges associated with the time and resource intensive tasks of analyzing textual data, as they apply to Engineering Education contexts, with automated techniques incorporating Natural Language Processing. This chapter first provides the necessary background for my work, which then motivates the research problem and methods. Finally, I present an overview of my research, and introduce the key contributions of my work.

1. Introduction

A report by the National Science Board (2016) indicated that in the United States, undergraduate enrollment increased from 13.3 million in 2000 to 17.7 million in 2013 (pg.5). To meet this growing enrollment, several universities have adopted large classrooms for courses (Parry, 2012), in order to maximize utilization of resources while managing increasing costs (Soledad & Grohs, 2017). Larger classes contribute to the surge in textual data since they yield large quantities of student-generated textual data, including homework, classwork assignments, and open-ended responses to questions on teaching evaluations. Often due to large volume, textual data can be difficult for instructors and administrators to process. For example, describing the data collected as

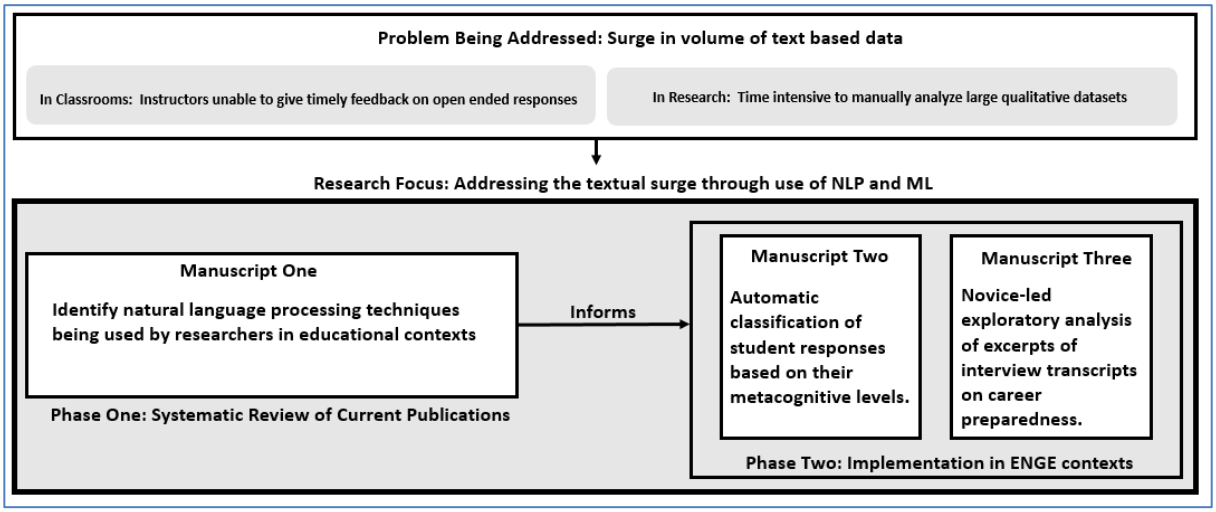
part of end-of-year course evaluations from students, Soledad and Grohs (2017) described how a lot of value exists in identifying experiences that students articulate as meaningful to the learning process through the student perceptions of teaching (SPOT) surveys. However, use of textual data such as those from the open-ended parts of course evaluations such as SPOT to inform pedagogy is not maximized (Blair & Noel, 2014). Failure to maximize information extraction through analysis of textual data may be attributed to the challenges of analyzing volumes of these textual data due to the time and resource intensiveness of the analysis task.

The increase in data over the years may lead to challenges such as those related to extracting meaningful information from large volumes of data for use by practitioners, administrators or researchers. Papamitsiou and Economides (2014) identify handling of large amounts of data manually as prohibitive, and acknowledge the growing need for sophisticated analytical techniques to evaluate datasets for education related contexts. Exploiting computational capabilities for automated analysis of datasets may be one such example of sophisticated analysis. In recent years, there has been a 'boom in the computational capabilities of machines' contributing to the increasing use of computer based analysis of datasets (Baker & Inventado, 2014). Natural Language Processing, which refers to use of computer manipulation of natural languages (Bird, Klein, & Loper, 2009) may be useful for automated textual analysis and can be successful in extracting meaningful data from large volumes of text, and thus help inform research, practice and pedagogy.

Identifying the usefulness of Natural Language Processing techniques for faster and automated analysis of large volumes of textual data, my research seeks to explore application of these techniques for data generated in Engineering Education contexts. Examples of textual data in Engineering Education include, but are not limited to, publications, interview transcripts, student responses to open ended questions on assignments, and survey responses. The purpose of my doctoral research was two pronged: (1) to describe the current state of use of Natural Language Processing as it applies to the broader field of Education, and (2) to demonstrate the use of Natural Language Processing techniques for two Engineering Education contexts- one classroom context and one research context.

I first conducted a systematic review to describe the current state of use of Natural Language Processing as it applies to the broad field of Education. I specifically looked at the Natural Language Processing techniques used in published research articles, and the context for the use of such techniques. The findings of my systematic review informed my demonstration of the use of Natural Language Processing for two Engineering Education (ENGE) specific contexts: (1) ENGE classrooms: automated classification of open ended reflective student responses to gauge metacognition levels in undergraduate engineering classrooms, and (2) ENGE research: using text analytics for a novice-led exploratory analysis of excerpts from student interviews related to career preparedness. For the first context, I developed an automated system using Natural Language Processing techniques for the task of classifying student responses to open ended prompts, in order to identify the extent to which the student is engaged in metacognitive

behaviors (e.g., high vs low metacognitive development) in a given course. This algorithm can help instructors understand the metacognitive development levels of students in their classrooms. For the second context, I have described how Natural Language Processing can be used for a novice-led exploratory research on qualitative



datasets. In this second application, I have established how Engineering Education researchers can efficiently engage with qualitative data generated from student interviews, through an emergent and iterative approach to analysis.

Ch1. Fig. 1: Overview of the Research Focus, detailing the problems being addressed

2. Background

A brief overview of Natural Language Processing techniques as described for use in education research will be helpful to understand why these techniques are relevant as we consider time and resource effective solutions of tackling analysis of qualitative datasets.

2.1 An overview of Natural Language Processing

Bird et al. (2009) used the term Natural Language Processing to describe any kind of computer-based manipulation of a natural language. They distinguished between natural languages such as English, German, or Hindi from what they connote as artificial languages such as those used in programming, and observe that the former have evolved over time, passing from generation to generation with rules that may be hard to pin down explicitly. Natural Language Processing can be understood as an interdisciplinary approach in which computers are used to perform useful tasks involving human language (Jurafsky & Martin, 2007). Popular Natural Language Processing tasks include: machine translation (eg. Google Translate translating a phrase from Hindi to English), and automatic speech recognition (eg. Siri on iPhones recognizing voice commands to "call home"), among others.

Natural Language Processing techniques have been used in conjunction with machine learning tools for varied content analysis related tasks. One of the definitions of machine learning was provided in the IBM Journal of Research and Development by Samuel (1959) who in describing machine learning using a game of checkers, stated that these types of studies are concerned with the "programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning." (pg.71). Machine learning has been identified as a viable option for the task of automated thematic content analysis for large corpora (Scharkow, 2013; Sebastiani, 2002). In their paper on the empirical evaluation of systems for online content

analysis using supervised learning, Scharkow (2013) reasoned that machine learning processes to automate content analysis is a promising endeavor because:

> "(1) it directly uses manually coded documents as training data, (2) is language and topic-agnostic, (3) can be used and evaluated in the same way as conventional analyses and (4) requires little to no extra effort because data collected by hand-coding can be used to quietly train and test a classifier in the background." (pg.18)

Supervised machine learning techniques, such as the one used by Scharkow (2013) for content analysis, rely on statistical tools to enable the machine to iteratively learn from the data. Supervised machine learning has been described by Kotsiantis, Zaharakis, and Pintelas (2007) as algorithms that "reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances" (pg.249).

Thus, central to supervised machine learning are the labels provided to the algorithm. The goal of the supervised machine learning system is to model the labels based on a set of predictors, which will fit an unlabeled instance (Kotsiantis et al., 2007). Specific to textual content analysis for example, supervised machine learning techniques of classification have been used with Natural Language Processing techniques for tasks such as classification of suicide notes into genuine or illicit (Pestian, Nasrallah, Matykiewicz, Bennett, & Leenaars, 2010), and for analyzing over 300 State of the Union addresses by Presidents of the United States of America to determine a timeline of trends in topics of national interest (Savoy, 2015).

In contexts related to education, manual codes developed by education researchers or instructors may be used to train the automation system for supervised machine learning tasks. For instance, in the study by Leeman-Munk, Wiebe, and Lester (2014), the

researchers use text analytics to automatically grade constructed responses from students on an assignment, thus contributing to real time formative assessment in the classroom. In this supervised machine learning based system, the authors describe the input to the system to comprise of a set of responses and grades assigned by the instructor for those responses. Once trained on such labeled data, the system can then move to unlabeled responses and automatically assign grades to those.

2.2 Application of Automated Text Analysis for Education Contexts

Analyzing textual data generated in classrooms may be of particular interest to researchers, but may also be extremely useful for instructors and administrators. One example of use of textual data is to inform learning analytics. Learning analytics may be understood in terms of the definition provided by Oblinger (2012), as that which pertains to "students and their learning behaviors, gathering data from course management and student information systems in order to improve student success". Slade and Prinsloo (2013) expanded upon this definition and described learning analytics as the "collection, analysis, use and appropriate dissemination of student-generated, actionable data with the purpose of creating appropriate cognitive, administrative, and effective supports for learners" (pg.4). Text may thus be leveraged to understand how students learn in a classroom. However, as previously described, analyzing these texts are time and resource intensive for individual instructors. Therefore, Natural Language Processing could be helpful to leverage datasets for education contexts.

In fact, research already shows the utility of Natural Language Processing techniques to education. Examples of studies include use of Natural Language Processing techniques for the purposes of assessment (e.g., Magliano and Graesser (2012) or to understand student vocabulary (e.g.,Variawa (2014), Variawa, McCahan, and Chignell (2013)). Using Natural Language Processing tools with supervised machine learning for a predictive analysis, for example, Robinson, Yeomans, Reich, Hulleman, and Gehlbach (2016) presented a model which could predict course completion in MOOCs based on word choices of students enrolled in the course. In a study conducted for an Engineering Education classroom, Variawa et al. (2013) demonstrated how an automated method could be used to develop course-specific vocabulary. The authors further describe how their model successfully identified domain-specific terms on engineering exams using a modified tf-idf approach to assess the frequency of words used. For example, the authors discuss finding terms frequently used by students on exams to largely be part of course-specific vocabulary (e.g., words like CRSS as abbreviated notation for critical resolved shear stress). They found that their approach was able to distinguish between course specific words from other words used by students on their exams. In this study the authors also justified the use of computational methods to analyze educational textual data sets due to ease of use by instructors and researchers, and enhanced efficacy of such automated methods in dealing with larger data sets as compared to individual human analysts.

3. Organization

As previously described, my research study includes three manuscripts. In Manuscript One, I present findings from a *systematic review* of existing publications on the use of Natural Language Processing in educational contexts. The findings described in Manuscript One have informed the research and techniques implemented for the specific tasks described in Manuscripts Two and Three. Both Manuscripts Two and Three demonstrate the use of Natural Language Processing for textual data generated in Engineering Education contexts. Manuscript Two is an *analysis of engineering student responses* using Natural Language Processing to automatically classify these responses based on metacognitive levels. Manuscript Three is an *exploratory novice-led analysis of excerpts from student interviews* to uncover interesting themes that may be important for detailed follow up qualitative analysis. Table 1 provides an overview of the intended research design and the specific research questions.

Ch 1. Table 1: Overview of the Research Purpose, Questions, and Outcomes for each Manuscript

Objective	Purpose	Data Source	Research Questions	Outcomes
<u>Manuscript One:</u> Systematic Review on the use of Natural Language Processing in Education	To describe the current state of use of Natural Language Processing as it applies to the broad field of Education.	Articles collected from various databases, subject to an inclusion criteria	RQ1. How is Natural Language Processing used in education related contexts? RQ2. What are the strengths of these automated text analysis methods, as documented by the authors of the publications? RQ3. What are the limitations of these automated text analysis methods, in terms of challenges as documented by the authors of the publications?	Description of the contexts for use of Natural Language Processing techniques including strengths and weaknesses. Overview of tasks that Natural Language Processing was used for, and the specific techniques that were used by the authors.

Objective	Purpose	Data Source	Research Questions	Outcomes
<u>Manuscript Two:</u> Identifying student metacognitive levels as indicated through open ended responses.	To use textual analytics to classify student responses as high, medium, and low for level of metacognitive development.	Student artefacts such as reflection assignments	What characteristics can an automated textual analytics system identify based on students' classrooms assignments that can facilitate instructor categorization of students' metacognitive performance	Automated classification system for student responses to identify which are most indicative of High, Medium, or Low metacognitive development.
<u>Manuscript Three:</u> Understanding Career Preparedness Among Engineering Students	To report on how insights from Natural Language Processing techniques were used to facilitate exploratory analysis of a large interview dataset by a novice researcher, which in turn contributed to a team of experts' understanding of the dataset	Excerpts from student interview transcripts related to career preparedness	What insights can be gained from using Natural Language Processing techniques for exploratory novice-led analysis to inform opportunities for future analysis for a team of researchers working on understanding engineering student choices related to career preparedness?	Directions for potential future in-depth qualitative research on the same dataset Overview of the dataset in terms of word frequencies, word clusters, and topic modeling.

In Manuscript One, I systematically reviewed publications that elaborated on the use of Natural Language Processing for Education related contexts. Specifically, in this manuscript I looked at the techniques implemented by researchers, the tasks the Natural Language Processing was used for, and the documented the strengths and weaknesses of the implementations, as described by the authors. Results of this manuscript informed the choice of techniques implemented in manuscripts Two and Three. The results of Manuscript One thus provided an idea of the current state of use of Natural Language Processing techniques for educational contexts.

Drawing on the findings from Manuscript One, the other two manuscripts of this study led to development of automated systems using a variety of techniques to accomplish

various tasks. For example, POS (Part of Speech) tagging is a technique in which each word in a text corpus is labeled for the Part of Speech they belong to (i.e., Noun, Verb, Adjective, Adverb, etc.). The tagging of each word to identify the associated part of speech can help establish relationships and trends in the dataset For example, in their study to explore trends in 300 State of the Union Addresses, Savoy (2015) use Part of Speech as a feature to categorize addresses by different Presidents. The analysis presented by Savoy (2015) indicated that Obama predominantly used verbs in his speeches, while Hoover used mainly nouns in speeches indicating a trend of action-oriented speech in the former and more oriented towards explaining situations such as the economic situation of the 1930s in the latter. For the Manuscript Two, I developed a system using machine learning algorithms which statistically used information such as part of speech usage frequency, length of sentence, etc. to classify the responses into levels based on metacognitive development.

In Manuscript Three, I conducted a novice-led exploratory analysis which used Natural Language Processing to cluster the word choices of students in the interview excerpts, graph word frequencies and model topics from the dataset Manuscript Three helped a team of Engineering Education researchers to engage with a large body of textual data, to understand trends which provided them with insights for future qualitative analyses.

For each of the manuscripts detailed, I documented not only the algorithms used to implement the tasks identified above, but also provided a detailed memo-ing of the

choices made by the researcher and the benefits of such a machine-based approach to analysis of textual information.

4. Implications for Research and Practice

The contributions of my work may be summarized as: (1) it expands on research at the intersection of contemporary research areas of Natural Language Processing, Learning Analytics and Educational Data Mining, (2) it addresses the call for new methodologies for emerging challenges in Engineering Education, (3) it demonstrates the applicability of an innovative methodology to analyze qualitative datasets through statistical methods beyond counting, and (4) it provides significant insights of relevance to multiple stake-holders in Engineering Education such as instructors, researchers, administrators, and students.

One of the prominent implications of this work to Engineering Education is its impact on existing research on learning analytics and educational data mining for Engineering Education, by introducing Natural Language Processing techniques as a method for analysis. Slade and Prinsloo (2013) described learning analytics as the "collection, analysis, use and appropriate dissemination of student-generated, actionable data with the purpose of creating appropriate cognitive, administrative, and effective supports for learners" (pg.4). Siemens and Long (2011) acknowledged the value that data analytics bring to higher education. They described learning analytics as essential to "penetrate the fog that has settled over much of higher education" (pg. 30), and described how analytics

can provide researchers, educators, students and administrators a foundation upon which to enact change.

In my research thus, I demonstrated an innovative approach to analysis of textual data in two Engineering Educational contexts. In the first context, I demonstrated use of Natural Language Processing to automate classification of student responses to open ended prompts, while in the second context I demonstrated the use of Natural Language Processing techniques for exploratory novice-led research of a qualitative dataset

This research also has direct implications for Engineering Education practice. Through my systematic review of existing publications detailing on the use of Natural Language Processing I have provided an overview of how these techniques have been used in Education contexts. This overview will be helpful for instructors and researchers to help them think about ways to process large amount of textual data that may provide a wealth of information, but traditionally take a long time to analyze manually.

The intended overall impact of this research is directed towards helping Engineering Education researchers recognize value in Natural Language Processing techniques for exploring and analyzing textual datasets. Use of such techniques can help researchers, instructors and administrators leverage textual data generated in our discipline and conduct faster analysis on them to gain deeper understanding of engineering contexts. Parry (2012) elaborated on how big data analytics is currently being leveraged by universities to help personalize feedback in increasingly larger classrooms across universities in the US.

My research demonstrates that Natural Language Processing techniques can indeed be exploited to enable instructors and researchers quickly and meaningfully analyze textual datasets in Engineering Education contexts. This move towards automated analysis of textual data may be increasingly relevant in our present world as we progress towards completely electronic data bases. Using automated intelligence to extract information from textual corpora may be a sustainable solution for tackling the increasingly large amounts of qualitative data.

MANUSCRIPT 1: A MIXED METHODS SYSTEMATIC REVIEW ON THE USE OF NATURAL LANGUAGE PROCESSING IN EDUCATIONAL CONTEXTS

1. Introduction

The increasing volume of textual data in Education has led to the use of automated techniques to process the data in ways that are time and resource efficient and effective. In recent years, there has been a 'boom in the computational capabilities of machines' contributing to the increasing use of computer based analysis of datasets (Baker & Inventado, 2014). In fact, Papamitsiou and Economides (2014), referred to educational contexts identify handling of large amounts of data manually as prohibitive, and acknowledged the growing need for sophisticated analytical techniques to evaluate data, exploit patterns within the data, and eventually aid in decision making processes. Automated text analytics may be beneficial for education researchers and instructors due to possibilities of automated grading of constructed responses (e.g., Leeman-Munk et al. (2014)), predicting course completion based on automated assessments of word/phrase choices in student responses (e.g.,Robinson et al. (2016)) or development of course specific vocabulary (e.g., Variawa et al. (2013)). Thus, there are multiple ways in which automated text analytics have been leveraged in Educational contexts. In particular, for

textual datasets in Education, automated analysis techniques using Natural Language Processing may provide a feasible alternative to manual analyses processes.

Natural Language Processing can be understood as the use of computer manipulation of natural languages (Bird et al., 2009). Natural Language Processing may be used to successfully extract meaningful data from large volumes of text, and thus help inform research, practice and pedagogy in education, in a time and resource effective way. Time and resource effective means of analyzing text-based data, such as those incorporating Natural Language Processing, may be useful in the field of Engineering Education, especially in current times which have seen a surge in volumes of data being generated in the classrooms due to the effect of increasing enrolments (National Science Board, 2016) on class sizes (Parry, 2012), as well as through increasing research studies as demonstrated by increasing publications in the field (Borrego, et al. (2014)).

Borrego, et al. (2014) describe how Engineering Education being still in its infancy, borrows heavily from established methods in other fields such as the broader field of Education. I offer this systematic review to serve research and practice in Engineering Education by providing an overall picture of existing literature on how Natural Language Processing techniques have been used in the broader field of Education. I will use my systematic review to critically summarize the current state of automated text analytics using Natural Language Processing based techniques for educational contexts. Borrego, Foster, and Froyd (2014) characterize systematic reviews as those which follow transparent, methodical, and reproducible procedures to establish an overall picture from a collection of studies. Petticrew and Roberts (2008) describe systematic reviews as those

which offer an orderly and unbiased way of making sense of large bodies of literature to answer specific questions about that literature (as cited in P. R. Brown, McCord, Matusovich, and Kajfez (2015)). An orderly and unbiased understanding of the use of Natural Language Processing techniques in the broader field of education, may be helpful to engineering educators who may consider using these techniques in analyzing textual datasets.

This systematic review addresses the gap in literature specific to dearth of studies synthesizing articles describing use of Natural Language Processing techniques in education contexts. Thus, while there exist reviews, both systematic and traditional, which address the broader topics of educational data mining or learning analytics which include big numeric datasets (e.g., Baker & Inventado (2014), Papamitsiou & Economides (2014)); as well as synthesis of articles on topics related to the use of Natural Language Processing techniques in fields such as in medicine or business (e.g., Cormack, 2008; Demner-Fushman, Chapman, & McDonald, 2009; Stanfill, Williams, Fenton, Jenders, & Hersh, 2010), there are limited (if any) systematic reviews addressing specifically the topic of analysis of textual data using Natural Language Processing techniques specifically for educational contexts. The purpose of this systematic review is thus to describe the current state of use of Natural Language Processing as it applies to the broad field of Education. This purpose is helpful for engineering educators since it led to identifying opportunities for future research and innovation in the field of Engineering Education. To this regard, the specific research questions driving this study are:

RQ1. How is Natural Language Processing used in education related contexts?

RQ2. What are the strengths of these automated text analysis methods, as documented by the authors of the publications?

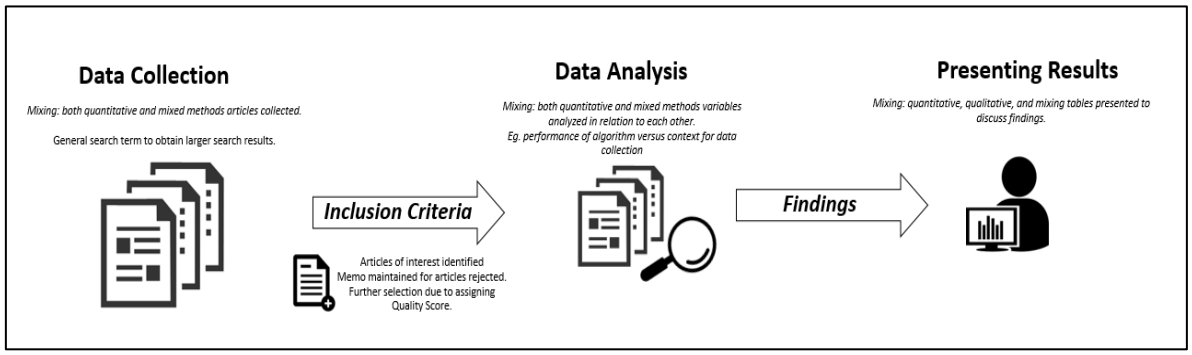
RQ3. What are the weaknesses/limitations of these automated text analysis methods, in terms of challenges as documented by the authors of the publications?

The answers to these research questions will help address my purpose which is to describe the current state of practice. Specifically, while answers to research RQ1 will provide an overall idea of the current state of use of Natural Language Processing techniques in Education contexts, the answers to RQ2 and RQ3 will help identify opportunities for future research, specifically in Engineering Education, by paying attention to the strengths and weaknesses/limitations of these methods.

For the purpose of this review, I borrowed the broad and intentionally ambiguous definition of Natural Language Processing as, ‘‘the use of computer manipulation of natural languages’’(Bird et al., 2009), and limit the contexts of the applications to those related to education research or instruction. Beyond the use of the term Natural Language Processing in the empirical studies, I attempted a grounded approach (Creswell, 2014), similar to Morelock (2017), in which my results are grounded in the views of the participants. In this systematic review, the participants were the authors of the papers included in the review. Thus, I followed an open and emergent approach to discovering themes across the articles based on analyzing the sentences that the authors had written to detail the use of techniques.

2. Methods

The steps followed by researchers while conducting a systematic review, typically begin with defining an inclusion criterion (Borrego et al., 2014; Borrego, Foster, & Froyd, 2015; Creamer, Simmons, & Yu, to be published; Evidence for Policy and practice Information and Co-ordinating Centre (EPPI-Centre), 2010; Petticrew & Roberts, 2008; Sandelowski, 2008). Once an inclusion criterion is defined, the articles included in the review are appraised to ensure relevance of the publications included in the review to the research purpose and research questions. The appraisal typically includes a quality score. In my review the appraisal score was assigned to the papers based on whether or not the article was relevant to the review based on the review questions, and if it the methods described were the methodologically transparent. Finally, the review is conducted systematically, by recording all the choices and decisions made by the reviewer to ensure high quality of the research.



Manuscript One. Fig. 1 Detailing the Steps in my Systematic Review Synthesis

Fig. 1 highlights the main steps in the review, and the following paragraphs elaborate on the inclusion criteria and quality score relevant to my study. However, before

providing further detail on the procedure for the systematic review, I will describe the larger methodological framework driving this syntheses. An understanding of the Mixed Methods approach to this synthesis, and the rationale for the choice of this methodological framework will prepare the reader for the following paragraphs related to the data collection, analyses and result interpretation.

2.1 Mixed Methods Research Syntheses

The EPPI Centre (2010) recommends using methods for the systematic review process that align with the research questions of interest. Petticrew and Roberts (2006) advise use of quantitative methods to answer review questions related to "what works?", and qualitative methods to answer review questions related to "what matters". My research synthesis is backed by an interest in summarizing both what works as well as what matters, in terms of how Natural Language Processing is used in Education contexts. This is because I seek to summarize the techniques used but also understand their contexts of use to determine the strengths and weaknesses/limitations of the approach. Thus, I conducted a systematic review using a mixed methods approach, often described in literature as mixed methods research synthesis (Sandelowski, Voils, & Barroso, 2006). Heyvaert, Maes, and Onghena (2013) state that while conducting systematic reviews, mixed syntheses, as compared to "un-mixed syntheses" may provide 'more complete, concrete, and nuanced answers' to complex research questions. Heyvaert, Maes, and Onghena (2013) further supply support for mixed methods research syntheses by stating that these syntheses are successfully able to answer multiple aspects of questions such as,

"what it is about this kind of intervention that works, and for whom, in what circumstances, in what respects, and why?" (pg.671).

Mixed methods research thus refers to the integration of the two approaches (qualitative and quantitative) in a research study. Greene (2008) describe mixed methods using a dialectical stance to understanding how different paradigms may engage dialogically with each other in mixed method research studies to generate new insights and understandings. Thus, key to mixed methods research may be the engagement and interaction of different approaches. Creswell and Plano Clark (2007) describe as the central premise of mixed methods research the integration of quantitative and qualitative approaches, which they believe provides a better understanding of research problems than either approach alone.

Mixing is integral to mixed methods, which are different from multi-method studies, which may use different methods, but fail to integrate them. One way to understand mixing is integrating quantitative and qualitative methods, data, to present a more holistic understanding of a phenomenon. In my study, I mix in multiple stages of the research. My mixed methods approach was concurrent (Creswell, Klassen, Plano Clark, & Smith, 2011) in that I collected and analyzed both quantitative and qualitative data together. In the data collection phase I did not restrict my inclusion criteria for either quantitative, qualitative, or mixed methods articles, rather my inclusion criteria did not screen for methods. In data collection there was mixing in terms of variables of interest identified in the primary articles which were both quantitative (e.g., date of publication) or qualitative (e.g., description of context, description of strengths and weaknesses

/limitations). In data analysis, I mixed by comparing across the quantitative and qualitative variables and mixing one another to determine a more holistic picture of automated text analytic use. Finally, in reporting the results of analysis, I included both descriptive tables (e.g., elaborating on strengths and weaknesses/limitations) as well as more quantitative graphs and plots (e.g., a timeline of studies reviewed).

2.2 Data Collection

For my review I followed steps described in exemplar articles describing systematic reviews conducted with mixed methods (e.g., Sandelowski, Leeman, Knafl, & Crandell, 2013; Sandelowski, Voils, & Barroso, 2006; Sandelowski, Voils, Barroso, & Lee, 2008; Voils, Sandelowski, Barroso, & Hasselblad, 2008). The first step in data collection was to identify an inclusion criteria. Using this inclusion criteria I then collected and critically appraised articles from multiple databases such as EBSCOhost: Academic Search Complete, Education Research Complete, ERIC, and ACM. I then created a spreadsheet for the articles in this review, and extracted demographic information related to these articles such as Year of Publication. The data was then used to answer my specific research questions.

2.2.1 Inclusion Criteria.

An inclusion criterion describes the type of primary articles included in the review, and is often directed by the purpose and research questions for the research syntheses (Borrego et al., 2014). My research synthesis included articles that provided insight to my

three research questions. Therefore, the article must have described how Natural Language Processing was used in a specific education related context, and additionally, may or may not have addressed (a) strengths of using the approach, and (b) limitations/weaknesses in the use of the techniques. Additionally, following in line with multiple systematic reviews conducted in the field of Engineering Education (e.g., Morelock, 2017; Wankat, 1999; Wankat, Williams & Neto, 2014; Whitin & Sheppard, 2004), I only considered articles that were from peer-reviewed sources (i.e., a journal publication or a conference proceeding), and which were available in English.

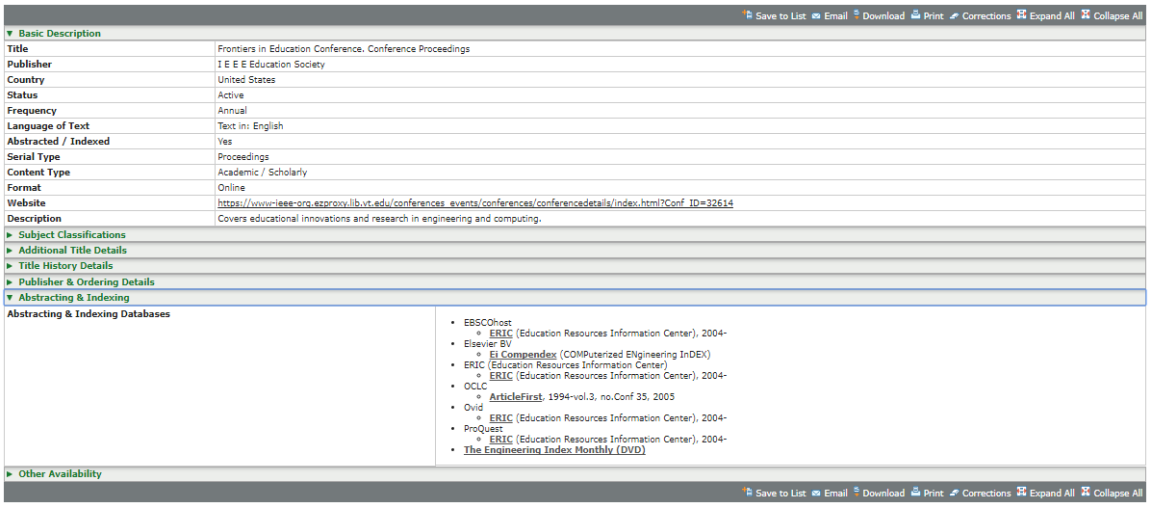
I restricted my timeline for the articles to the present decade (i.e., all studies published in 2010 and since). The articles considered for this review were restricted to the present decade intentionally to serve two purposes. The first purpose was to maintain relevance in terms of currency of techniques being discussed. The second purpose was to ensure that the timeline also made sense from a historical perspective of the larger field of educational data mining, due to the acknowledged growth of interest in Educational Data Mining (EDM) since late 2000s (e.g., 1st conference in EDM in 2008, Journal of Educational Data Mining first published in 2009 (Baker & Inventado, 2014)).

2.2.2 Scoping Study, Databases, and Search Terms

As recommended by Tranfield, Denyer, and Smart (2003) and implemented in the systematic review conducted more recently by Morelock (2017), a scoping study was conducted in April 2017 to test a preliminary set of databases and search terms, and to survey the breadth of literature surrounding Natural Language Processing for educational

contexts. Based on recommendations of McGowan & Sampson (2005) which was supported by Borrego et al. (2014), I consulted subject specialist librarians for Computer Sciences and Education / Engineering Education as well as a General Librarian for expert assistance in the search process. Based on the collective recommendations I searched the following EBSCOhost databases: Academic Search Complete, Education Research Complete, ERIC, Computers and Applied Sciences Complete, and Communication and Mass Media; as well as the Association for Computing Machinery (ACM) database using the following search terms: “Natural Language Processing” AND “Education”. The ACM database was added specifically on the recommendation by the Computer Science subject specialist librarian to ensure that articles from computer science related publication venues also found adequate representation in my systematic review.

In my consultations with the subject expert librarian for Education / Engineering Education, I discussed how it was important that the databases I chose indexed some publication sources in Engineering Education. The librarian helped me use Ulrich, to find



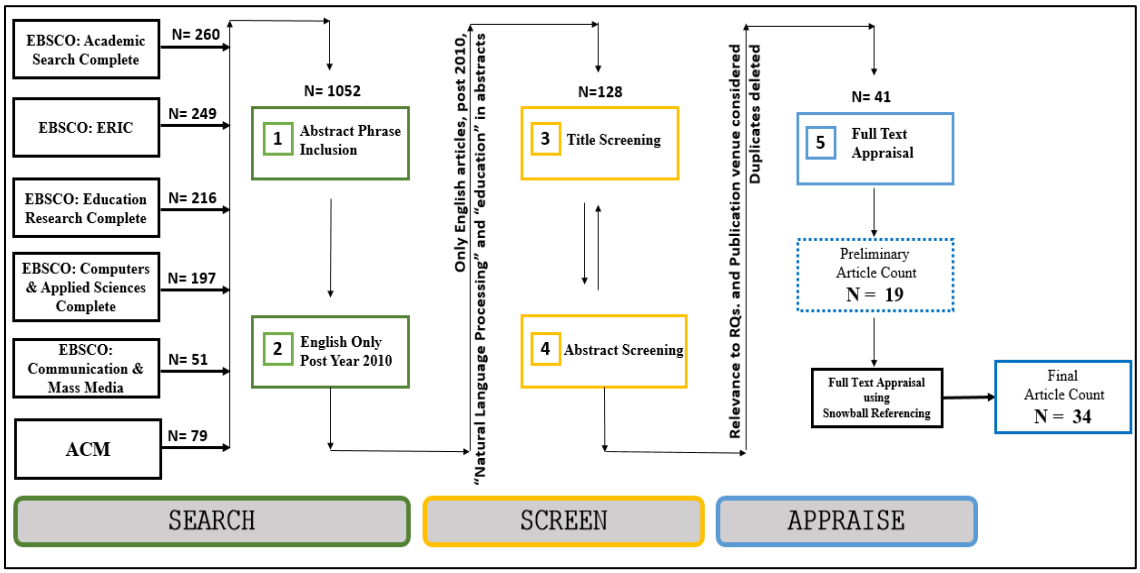
Manuscript One. Fig. 2: Ulrich Descriptor for FIE showing indexing by ERIC

out if specific journals were part of databases. Figure 2 shows the Ulrich description for how the online proceedings of the FIE Annual Conferences are indexed by ERIC.

Using Ulrich, I noted that articles from major Engineering Education sources, popularly used in systematic reviews in the field (e.g., Wankat (1999, 2004); Whitin and Sheppard (2004)) such as: Journal of Engineering Education (indexed in EBSCO Education Research Complete), International Journal of Engineering Education (indexed in EBSCO Education Source), European Journal of Engineering Education (indexed in EBSCO Education Source), and conference proceedings from the Annual Conferences of the American Society for Engineering Education (indexed in EBSCO Education Source) and Frontiers in Education (FIE)Annual Conferences (indexed in ERIC) were indexed by databases part of the EBSCOhost vendor. Thus a wide collection of databases were searched as advised in literature (e.g., Borrego, et al., 2014) and by subject expert librarians, to ensure that relevant studies were located despite the discipline of the journal the study was published in, or the database the study was indexed in.

2.2.3 Results and Filtering

I used an adaptation of the Search-Screen-Appraise methodology advocated by Borrego, et al. (2014). Figure 3, adapted from Borrego, et al. (2014) provides a visual representation of the steps in my data collection process, and the number of articles filtered at each step.



Manuscript One. Fig. 3: Five steps in the processes of Search, Screen and Appraise

As shown, there were five steps in this process. In the first step, I first found the results via database searching, and then filtered the search results based on the inclusion criteria that both terms "Natural Language Processing" and "education" should be part of the abstracts of the articles included. In the second step, which was also a part of Search, I restricted language of articles considered to English, published after 2010. The remaining articles were considered for steps 3 and 4 to screen the articles. The third step was to manually read the titles and assess if the article was relevant to the review. In this step, I also filtered out articles which were not part of conference proceedings or journal

sources. For example, the study by Skinner (2015) seemed relevant to the research questions, but was part of a doctoral dissertation, and thus not included in this review, as per my inclusion criteria of restricting articles to peer reviewed conference and journal publication venues. In case of any ambiguity in terms of relevance of an article to the review, the article progressed to the fourth step of abstract screening, leading to multiple iterations between steps three and four. In the fourth step, I manually read the individual abstracts and excluded those that seemed irrelevant to the scope of this review. For example, the paper by Zhu, Yan, and Song (2016) had both the terms "Natural Language Processing" as well as "education" in the abstract, however, this was because the paper described the evolving academic landscape through faculty hiring data, and these were specializations of the faculty. This article, and similar work irrelevant to the scope of this review were excluded in this step. Articles with incomplete or ambiguous abstracts, however, made it to the next step.

Finally, in the fifth step, I conducted Appraisal through a full-text analysis of the articles that had passed step four. For each article I assigned a quality score based on whether the article was relevant to my research questions (Relevant = 1, Not =0) and whether or not the authors followed methodological transparency in describing their study (Methodologically transparent = 1, Not = 0). Thus, articles could attain scores of 0 through 2. Articles with scores less than 2 were discarded (i.e., were either not relevant, or were not methodologically transparent). After the full text appraisal, only studies that met all inclusion criteria were included as part of the review. In terms of relevance, for example, only studies that described how Natural Language Processing techniques were

implemented in specific education contexts were included. This led to exclusion of studies like the one by Azam (2011) which addressed the scope of Natural Language Processing in education, but was a literature review on the issues and challenges related to speech generation by intelligent systems. Similarly, in terms of methodological transparency, it was important that the articles had methodological transparency while describing the Natural Language Processing techniques. If the articles did not elaborate on how or what techniques were used, but rather provided an example of a black-box software, it was excluded from the review. An example is the article by Zeng-Treitler et al. (2014) in which the authors describe the use of a computer application called Glyph to convert text to set of illustrations to provide pictures for health communication. While the context was described adequately, as were the evaluation of the results, because there were no details provided for the Natural Language Processing techniques themselves, this study was excluded from my review.

After full text appraisal, I found that I was left with 19 articles for inclusion in my review. I then conducted snowball referencing (Borrego et al., 2014) on the 19 articles in my dataset, and included 15 more articles which were referenced by the original 19 articles found through the database Search-Screen-Appraise process. All of these 15 additional articles passed all my inclusion criteria through full text appraisal. Thus, a total of 34 articles were finalized for inclusion in my review.

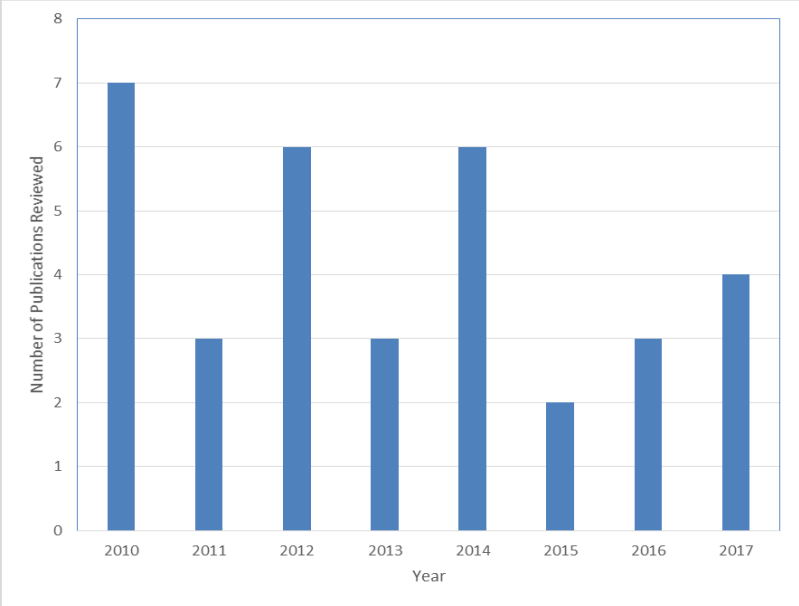
2.2.4 Demographics of Included Publications

The articles included in the syntheses were coded for demographic information related to the year of publication, the source of the article (i.e., conference or journal), and

whether or not it was used in an Engineering Education context. An understanding of these demographic details for the articles included in the review can help us think about the interest among researchers on this topic (as evidenced through a consistency in the number of publications per year) as well as provide those who may be interested in exploring similar techniques with an idea about the venues these are published in (i.e., predominantly conferences as opposed to journals).

2.2.4.1 Year of publications. As mentioned in an earlier section, I had enforced a date restriction to only include studies published since 2010 for this synthesis.

Figure 4 shows the graph of the number of publications for each year in the eight years



Manuscript One. Fig. 4: Depicting number of articles per year considered

included in this review. Note that since all the publications from 2017 will probably not be indexed until early 2018, the numbers for 2017 are incomplete, and hence not insightful. We can see from the data that the number of publications were more or less consistent across the years, which can be interpreted as a continued interest among

researchers for this topic. The interest among researchers in Natural Language Processing to its application specific to Education is also demonstrated by the fact that there exists a series of workshops (now in its twelfth year) devoted specifically to Building Educational Applications (BEA) using Natural Language Processing. These workshops are the Annual Workshop on Proceedings of the Association for Computational Linguistics' workshop on Building Educational Applications Using Natural Language Processing, multiple papers from which were included in my review.

2.2.4.2 Source of publications

Based on the source of the publications, the articles were classified as either belonging to Conferences or to Journal publishing venues. In my dataset, close to 62% of the data (21 out of 34) were from Conference venues, while the rest were from Journal sources. The high percentage of articles from conference venues may indicate that these sources are the primary places where research related to Natural Language Processing applications specific to education are being published.

2.2.4.3 Engineering Specific contexts

During my review of the articles, I created a binary variable 'Engineering Education relation' to indicate whether or not the study was related to Engineering Education. Out of the 34 publications considered, 11 were related to Engineering contexts based on the datasets that were used for the studies. Thus, if a study was conducted for an engineering related population (engineering school, engineering class) the variable received a value =1, otherwise the default value for the Engineering Education related variable was set at 0. Thus close to 30% of the articles sampled related to Engineering

Education, although the authors themselves did not tag these articles as such, nor were the venues of publications related to Engineering Education for most of these publications.

2.3 Tracking and Data Analysis

I recorded a detailed description, including the demographics for each article, along with characteristics correlated to my three research questions in an Excel spreadsheet All of these variables were created deductively based on my research question and purpose to form a broad overview describe the current state of use of Natural Language Processing in Education. Thus, I was interested in how Natural Language Processing was used in terms of the purposes (e.g., providing students with automated grades) that were addressed through use of Natural Language Processing, and the specific Natural Language Processing tasks and techniques (e.g., development of an automatic classifier to classify student responses into pass or fail) that were used to address these education related purposes. I was also interested in the context of implementation (e.g., online MOOC based classrooms), and additionally extracted information relating to strengths and limitations/weaknesses of the approaches as described in the articles. Specifically, for each article, apart from demographic variables for year of publication, source of publication, keywords for the research, information on the following variables were also noted: Purpose, Specific Natural Language Processing Tasks and Techniques, Overview of the Contexts, and Strengths and Limitations.

2.3.1 Purpose

I coded the articles to indicate the broad education purpose for which the Natural Language Processing based system was used. In my analysis, I found that the purposes for the use of Natural Language Processing techniques could be categorized into four major categories: related to creating educational content, related to automated instruction or tutoring, related to assessment, and finally those that were more broad research topics beyond the scope of instruction or teaching. The findings from this classification are detailed in the Findings section and specifically show in Table 1 (pg.37) provides an overview of the purposes for which Natural Language Processing was used for in the articles included in part of the synthesis. I also provide an example article for each type of purpose.

2.3.2 Specific NLP Task and Techniques

I coded the articles to indicate the specific Natural Language Processing tasks and techniques used. Note that classification is the primary task that was conducted with close to 60% of the articles describing use of Natural Language Processing for classification purposes. To conduct these tasks, the authors used a variety of techniques. Some techniques were based on word frequencies such as tf-idf (Bhaduri & Roy, 2017; Variawa et al., 2013). In other cases, articles also elaborated on the use of techniques to cluster the words automatically based on semantic similarities of the words in the dataset (e.g., Mora, Ferrández, Gil, and Peral (2017)). In case of multiple techniques all techniques were coded individually. The findings are detailed in the Findings section and

specifically show in Table 2 (pg.38) provides an overview of the Natural Language Processing techniques which were used in the articles included in part of the synthesis.

2.3.3. Overview of the Context of Implementation

My next set of variables captured the details of the dataset for the implementation of the Natural Language Processing techniques. I created a quantitative variable, Dataset Size, to indicate dataset size for each article. The datasets for the implementations varied largely. This variable was supported by a qualitative variable describing the context for data collection and analysis. Most authors explicitly detailed the number of training and testing data points in their dataset For example, in using a classification system on sentences, Ozturk, Cicek, and Ergul (2017) explain how their datasets comprised 4652 tweets related to the courses. In the study by Feng, Saricaoglu, and Chukharev-Hudilainen (2016) the authors describe their large dataset comprising 55,000 sentences sampled from essay drafts collected from students.

2.3.4. Strengths and Limitations

From the articles, I extracted the sentences used to describe the strengths of using Natural Language Processing based approaches for the particular context, as well as those used to describe the strengths and limitations of using Natural Language Processing based approaches for particular context. These were then coded, and found to belong to two broad categories. The first category was for general strengths and limitations/weaknesses of use of Natural Language Processing in Education contexts. The second category was more specific and related to the specific study context or approach being used.

As can be seen, the variables are both quantitative (e.g., Dataset Size, Year of Publication) and qualitative (Context Description). I followed an open and emergent coding (Creswell, 2014) technique for the qualitative variables for context, strengths and limitations, which led to code categories for each of those three variables.

2.4 Quality of synthesis

Gough, Oliver, and Thomas (2012) reason that since systematic reviews are like any form of research, they can be conducted well or poorly, and thus, appropriate quality assurance processes are needed to evaluate them. Quality of systematic reviews can be understood in terms of the review process itself, as well as the quality of the primary studies included in the review. I describe two broad set of ways in which this synthesis has been ensured to be of high quality: (a) quality as a result of primary studies included in this syntheses, and (b) quality as a result of how the review was conducted and presented.

2.4.1 Quality as a result of primary studies included in this syntheses

Davies (2000) argued that the quality of systematic reviews depends not only on the rigor, transparency, and reporting of the inclusion and exclusion criteria by the reviewer, but more importantly on the quality of the primary studies on which the review is conducted. To ensure a high quality of the review as a result of the primary studies included in the synthesis I ensured steps to reduce publication bias (Davies, 2000), followed explicit and detailed inclusion and exclusion criteria, and finally conducted

thorough critical appraisal of the included articles. Borrego et al. (2014) state that the quality and thereby, impact of systematic reviews, are improved by inclusion of a broader body of literature. Recommending ways to reduce publication bias, i.e., bias as a result of the primary articles included or excluded in a review, Davies (2000) suggests that systematic studies need to include "extensive, if not exhaustive" searches for articles in different databases, conference proceedings, textbooks, journals, etc. Borrego et al. (2014) suggest reviewers to actively seek out and consider counterexamples to avoid selective review of evidence, including thorough searches of multiple database and gray literature. It is necessary to have explicit inclusion criteria, to ensure quality of the results of the systematic review and to minimize bias. Borrego, et al. (2014) further recommend that authors provide explicit details on why studies were dropped from systematic reviews, in addition to the inclusion criteria.

In my study, I have explicitly mentioned the rationale behind my inclusion and exclusion criteria. I have also intentionally included a wide range of publication venues including conferences and journals (although I excluded grey literature, rationale for which is presented in the section addressing limitations). In addition to the wide scope of publication venues I ensured a wide database for collection of the articles by including all subject related databases under EBSCOhost. Finally, systematic critical appraisal of studies included in the review is also recommended in the report by the Evidence for Policy and practice Information and Coordinating Centre (EPPI-Centre) (2010). I have described in an earlier section how I conducted critical appraisal of the included articles

by manually going through each article to ensure that the article contributed to the review by being able to answer my research questions.

2.4.2 Quality as a result of how the review was conducted and presented

Apart from factors that influence the quality of the review resulting from the primary studies included in the synthesis, there are other criteria to gauge quality of systematic reviews based on how the review is conducted and reported. For my systematic review, I supported the trustworthiness while conducting my synthesis in the review process through researcher triangulation of codes (Leydens, Moskal, & Pavelich, 2004), as well as maintained transparency (Hiles, 2008) in the dissemination of the findings.

Trustworthiness of systematic reviews may be enhanced through supporting the reliability of the steps in the research process. Similar to Morelock (2017), to support the trustworthiness of my review I conducted researcher triangulation through an inter-coder reliability process (Leydens, Moskal, and Pavelich, 2004; Merriam, 1995 as cited in Morelock (2016)). I randomly chose twenty articles from my list of articles selected for review, and presented ten each to two colleagues along with a list of my codes and their corresponding definitions. Both of these colleagues were graduate students in the Department of Electrical and Computer Engineering, studying machine learning with knowledge of Natural Language Processing techniques. I requested each of these two colleagues to apply codes that they deemed appropriate for the ten documents assigned to them. I then met with both of them individually to compare my set of codes to their codes

for the same articles. For both sets of articles, we were able to reach an agreement upon the appropriateness for the codes that were applied by me initially. In situations where we disagreed on application of a particular code, I worked on revising the definitions of the codes until agreement was reached.

Transparency may be considered as another important criterion to gauge the quality of systematic reviews. Transparency in presentation and dissemination of findings is identified by Hiles (2008) as a benchmark for writing up research, who described transparency as the need to be explicit, clear and open about all the methods and procedures used in conducting research. Transparency can thus be understood as explicit detailing of the rationale for use of methods, stating research questions, detailing inclusion and exclusion criteria, and in reporting analysis and results in a methodic, accountable way which may be replicable. O'Cathain, Murphy, and Nicholl (2008) further list describing the justification for using a mixed methods approach to the research question; and describing the design in terms of the purpose, priority and sequence of methods as important criteria exemplifying good reporting of research studies (specifically mixed methods). In this write-up describing the review, I have maintained a high degree of transparency in reporting the results of my study, especially the rationale for choice of methods, the inclusion and exclusion of articles, as well as the steps in the analysis. Appendix One lists all included papers studied by this review.

2.5 Limitations

Although this review attempts to comprehensively ascertain the breadth of use of Natural Language Processing techniques in education, there are two sources of bias that limit this review. First related to the inclusion of the primary articles, I included only reviews written in English and those published in journals or conference proceedings. As a result, dissertations, non-academic reports, and grey literature were excluded, which may have yielded contributory insights to this review. Borrego et al. (2014) note that while grey literature may provide original empirical data and conclusions, they are an important component of the more traditional narrative reviews and are less relevant to systematic reviews seeking to understand specific interventions. While my review does not seek to understand a specific intervention, it does seek to understand the current state of use of Natural Language Processing techniques, and thus it seemed reasonable to exclude articles from grey literature sources from the scope of this review.

The second limitation of this review is that no thesaurus terms for Natural Language Processing or Education were used, thus excluding automated text analytical processes which may have used other terms to describe techniques under the broad umbrella of text analytics in education. However, given the large number of peer-reviewed articles that were included as part of the current review, it may be practical to assume that the current synthesis, although not exhaustive, presents an extensive overview of the scope and extent of use of Natural Language Processing techniques in education.

3. Findings

The main findings of this study include the wide range of purposes for which Natural Language Processing was used, as well as a list of the techniques used for these purposes. A wide range of contexts from online forums to in-class conversation feedback was also observed. A majority of the authors described the benefits of automation, but identified as a challenge the low volumes of annotated datasets, as well as lack of semantic information or context for the systems.

3.1 How is Natural Language Processing used in education related contexts?

Analyzing the articles included in the review, led to an understanding of four main purposes that authors described for using Natural Language Processing. Table 1 shows the purposes along with examples of articles. The purposes were predominantly related to automating the instruction process, through creation of course content, automating feedback or assessment, and creating automated tutors to help students as they learned concepts.

Manuscript One. Table. 1: Depicting the main purposes for use of NLP in Education

Purpose	Example of Article
Related to Creating Educational Content	E.g., Jayakodi, Bandara, Perrera and Meedeniya (2016) classified exam questions based on Bloom's Taxonomy to assess course content
Related to Instruction or Tutoring	E.g., Feng, Saricaoglu and Chukharev-Hudilainen (2016) described automatic error correction to help second language speakers of English develop grammar proficiency.

Related to Assessment	E.g., Basu, Jacobs and Vanderwende (2013) explained how a clustering approach could be used for automating short answer grading.
Not related to Instruction/Exploratory	E.g., Bhaduri and Roy (2017) demonstrated the use of Natural Language Processing to compare mission statements of engineering colleges from private versus public universities.

The purposes described by the authors drove the choice of Natural Language Processing techniques implemented. In my research I found that a majority of the studies presented findings related to the Natural Language Processing techniques for automated classification, this was followed by studies elaborating on clustering techniques. Other tasks included pattern recognition, error detection, and text adaptation. The Table 2 below provides an overview of the different types of Natural Language Processing tasks and the corresponding techniques that were found in the articles analyzed.

Manuscript One. Table. 2: NLP Tasks, and Techniques along with examples for same

NLP Task	Examples of techniques implemented
Classification	Ozturk, et al. (2017) explain how a naïve bayes classifier was used to classify tweets related to instruction on an open and distance education university into three sentiment classes: positive, negative, and neutral. Ramachandran, et al. (2016) used multi-class SVMs (Support Vector Machines) and logistic regression classifiers to classify reviews as either those detecting which were positive, negative, or advisory. Jayakodi, et al. (2014) used WordNet and Cosine Similarity to classify exam questions using Bloom's taxonomy.
Clustering	To understand what the most relevant topics posted as part of comments on forums of MOOCs were, Mora, et al. (2017) conducted hierarchical dendogram clustering to find topics of relevance. Xiong, et al. (2012) described how clustering tools were used to help instructors discover patterns that are reflected in peer reviews. Lan, et al. (2015) explain how clustering was used to automatically assign grades to solutions for mathematical assignments.
Pattern Recognition	

	Using word-order graphs, together with semantic relatedness mtrics, content of a review was easily identified (Ramachandran, et al., 2016).
Error Detection	Hui-Hsien Feng, et al. (2016) used a hybrid approach to error detection to enable grammatical error detection of students' academic writing.
Text Adaptation	Liu, et al. (2010) elaborate on how they used Natural Language Processing techniques at lexical, phonological, syntactical and semantic levels to design a computer assisted item authoring (text creation) of items for testing of elementary Chinese. Watanabe, et al. (2010) elaborate on how text simplification and summarization techniques using linguistic processing, knowledge of parts-of speech tags, can be achieved for helping low literacy readers access online information more easily. Dzikovska, et al. (2014) explain how Natural Language Processing techniques were used to generate automatic feedback for students, thus, leading to an intelligent tutoring system.

The contexts of the implementations for the articles considered as part of this review also varied greatly. The datasets themselves included a wide range of qualitative dataset from tweets (e.g., Ozturk, et al., 2017) to transcripts of recorded conversations (e.g., Shaw, et al., 2014), mission statements of colleges (e.g., Bhaduri & Roy, 2017) to exam questions collected from faculty (e.g., Jayakodi, et al., 2014).

3.2 What are the strengths of these automated text analysis methods, as documented by the authors of the publications?

I open coded the sentences extracted from the articles to understand the strengths described by the authors. I found codes for strengths related to general Natural Language Processing use in Education, and specific use for the specific studies. 29 of the articles out of the 34 articles reviewed explicitly described the benefits of Natural Language Processing in analyzing large datasets. These strengths were related to codes such as 'less

time consuming', and 'automated decision making'. For example, Trausan-Matu, Dascalu, and Mihai (2014) elaborated on how use of Natural Language Processing techniques helped in a 35% reduction in time for providing feedback for the instructor.

Manuscript One. Table. 3: Tabulating the Strengths of NLP approaches as described by the authors

Code Category	Open Codes	Examples from Articles
General	Related to Automation	"managers from the institution can concentrate on the shortcomings of the system and student complains by using outcome of the study" (Ozturk, et al, 2017) Mora, et al. (2017) describe how automated processes are less demanding and not time consuming as can be used to look for information about students.
Specific	Anonymized feedback	Students enrolled in the Anadolu university are not able to share their feelings, opinions, and complaints freely. Ozturk, et al. (2017) claim that using Natural Language Processing may be the best way to reach the students' real opinions.
	Less time consuming	"Both students and teachers may not have time to spend on reading all posts generated each week by users in the forums" Mora, et al. (2017) Xiong, et al. (2012) explain how instructors they interviewed complained that peer reviews were time consuming to read and impossible to interpret Natural Language Processing systems were

	useful to provide the instructors with a time-efficient way of engaging with the peer reviews.
Automated Decision making	Forums can be re-organized by the machine based on topic of relevance of each post on the forums, thus making it easier to track information for both students as well as teachers (Shaw, et al., 2014).

3.3 What are the limitations/weaknesses of these automated text analysis methods, as documented by the authors of the publications?

I open coded the sentences extracted from the articles to understand the limitations/weaknesses of the techniques as described by the authors. I found that not all articles discussed the limitations of their approach. This was especially true for conference publications which did not go into complete methodological transparency in detailing the methods. Some of the authors provided some discussion on the specific limitations of their implementations. For example, describing challenges, Oztuk, Cicek and Ergul (2017) in addition to commenting on more general limitations such as low sample size for training and testing also elaborated on challenges specific to their dataset related to the lack of a Turkish language based sentiment lexical.

Manuscript One. Table. 4: Tabulating the Limitations/Weaknesses of NLP approaches as described by the authors

Code Category	Open Codes	Examples from Articles
General	Inability to support certain data types	Jayakodi, et al. (2015) explained how the Natural Language Processing system was unable to process information from images, which was one of the limitations of their implementation to classify exam questions based on Bloom's taxonomy.
	Time and large number of options for tools	Shaw, et al. (2014) explained how they only used default values for the parameters for their techniques due to constraints of time.

Specific	Limited data (in terms of size of training and testing sets)	"Limited data provides as average success rate" (Ozturk, et al., 2017) in describing the results of their specific sentiment analysis task.
	Lack of lexicon models or context information to support analysis	"Turkish sentiment lexicon did not exist" (Ozturk, et al., 2017) Mora, et al. (2012) explained how certain keywords could be incorrectly categorized due to lack of context and knowledge of semantics in their system.

4. Discussion

4.1 Discussion Related to Findings of the Research Synthesis

Baker & Inventado (2014) attributed four factors leading to the emergence of using analytics for educational datasets: a substantial increase in the data quantity, improvements in the data formats, advances in computing, and increased sophistication of the tools available. These four factors may have contributed specifically to the increasing popularity of sophisticated techniques for Natural Language Processing related tasks. My review synthesis shows that there is an opportunity for use of Natural Language Processing for a wide range of purposes in Education. I have elaborated on the different techniques that studies have demonstrated using, as well as the strengths and challenges of using Natural Language Processing based approaches.

Researchers interested in implementing such techniques will have to search for them beyond disciplinary silos. This is because one of the observations from my dataset was that the venue for presentations of the findings were predominantly conferences that were related to computer science audiences. Only about 40% of the dataset was from journal

sources. This finding also highlights that educators interested in making their techniques more visible to the education community would need to start presenting their findings in education related conferences with larger attendances from practitioners who will be more directly impacted by this work.

Another pertinent observation is the apparent lack of a community of scholars who seem to be working on these techniques. A community of scholars is likely to develop with more advancement and interest in this research area. A strong and networked community of scholars would likely help those interested in applying these methods to leverage methodological guidance from the research experiences of those who have prior work experience in these processes. This exchange of results and best practices would then likely strengthen the research processes in this area.

This systematic review captures collaborations across institutions and countries (e.g., co-authors from U.K. and U.S.A), and a wide range of countries (e.g., authors who identified themselves as part of institutions in Turkey, Egypt, Romania, Canada, Ecuador). There is only one instance in which two papers are co-authored by the same set of authors (Xiong, and Litman on automated peer reviews). However, there doesn't seem to be any other overlaps between the authors or any explicit network connecting the authors. This may suggest that the authors may be working on their projects in isolation. That the authors are working in isolation, and that there seems to be an apparent lack of a scholarly community or network, is also evidenced by the scattered publication venues. Apart from the Building Educational Applications using NLP (an annual conference, hosted primarily in the United States), there doesn't seem to be any single publication

source or venue for the authors to publish their findings. These observations indicate that although there is an observed interest in using Natural Language Processing for educational contexts, there is definitely not a widespread use of these techniques.

4.2 Implications for Practice and Research in Engineering Education

My research has direct implications for Engineering Education research and practice by describing the current extent of use of Natural Language Processing techniques in education, which can also be used in similar engineering education contexts. I have shown through data support for literature on how in current years, automated text analytics have maintained consistent presence in Education as tools to understand and extract information from large volumes of data (Baker & Inventado (2014), Papamitsiou & Economides (2014)). While there are multiple software packages (R, SPSS, JMP), in place to make meaning from numeric data in time effective ways, textual data, unlike numeric data, is often more tedious and time consuming to analyze. This is especially true for the engineering classrooms. Consider as example, end of semester course evaluations for an engineering class. These evaluations may be collected using surveys administered to the students, which typically have a mix of Likert type questions (e.g., Please rate your instructor in terms of creating an inclusive learning environment) to which students may provide numeric responses, and some open-ended prompts to which students provide reflective text based responses.

While there may be a wealth of information in the open-ended articulations that students provide, it has been seen that often these open-ended responses are neglected in analysis, and serve as missed opportunities to gain deeper insights into the students' learning process (Soledad et al., 2017). This is usually due to the large class sizes in engineering, making it difficult for instructors or evaluators to manually assess individual written responses. Thus, there is a lack of maximizing information extraction through analysis of text-based data to inform pedagogy (Blair & Noel, 2014), and this may perhaps be attributed to the challenges of analyzing large volumes of these text-based data manually due to time and resource intensiveness of the analysis task (Soledad et al., 2017). In such scenarios in engineering classrooms, Natural Language Processing techniques may be useful to analyze text-based data in time and resource effective ways.

5. Conclusions

This mixed methods systematic review acknowledges the usefulness of Natural Language Processing techniques for education research and practice, and presents a synthesis of articles describing use of Natural Language Processing techniques for educational contexts. Thus the first and most significant contribution of this work is in terms of an overview of the different ways in which Natural Language Processing techniques have been used in educational contexts. It is hoped that engineering educators will be able to use the findings of this review to inform their practice or research methods to incorporate more automated analysis and facilitate their analyses of textual data to

yield time and resource effective insights, in view of the textual surge in engineering education both in the classrooms and in research.

I have used a mixed methods approach to conducting this systematic review, because I was interested in not only how the techniques were used but also why the techniques were used in terms of the strengths and weaknesses as documented. Mixed methods research syntheses are new to Engineering Education, and thus, a second contribution of this work in terms of the methodological transparency to this synthesis which may be useful to those considering conducting mixed methods systematic reviews of their own.

MANUSCRIPT 2: NLP IN THE ENGE CLASSROOMS USING AUTOMATIC TEXT CLASSIFICATION IN GAUGING METACOGNITIVE LEVELS

1. Introduction

Despite the many benefits of helping students develop metacognitive skills (Pintrich, 2002; Veenman, Van Hout-Wolters, & Afflerbach, 2006), it has been found that gauging and thus facilitating student metacognitive development can be time and resource intensive. The challenges of facilitating student metacognitive development may be attributed primarily to metacognition itself being a difficult construct to measure (McCord & Matusovich, 2013). Metacognitive development has been found to be effectively studied through analysis of reflective prompts; since metacognition essentially involves knowing about and regulating one's own cognitive processes (A. L. Brown, 1997; Flavell, 1979), and because the thinking happens internally. Assessing and giving feedback on reflective prompts however may lead to time and resource constraints, especially in large engineering classrooms. Cuseo (2007) elaborate how there is often limited opportunity for instructors to work with and give students feedback, given the increasing class sizes in universities.

In order to help instructors assess reflective student responses, this study reports on the results from development and use of a Natural Language Processing based automated text

classification system which can be used by instructors in engineering classrooms to give meaningful feedback to their students on the development of the students' metacognitive skills; as well as for the instructors themselves to understand the overall level of metacognitive development in their classrooms. In my study I analyzed student responses generated from reflective prompts collected as part of an intervention in three level engineering classes in the United States, to gauge student metacognitive development. Thus, my research focused on the development and use of Machine Learning and Natural Language Processing based text classification to mimic the metacognition expert, and categorize student responses into high, medium, and low based on the choice of words/phrases in the submitted assignments. The specific research question directing this study was:

RQ: How can Natural Language Processing be used to automate the process of classifying student responses to reflective prompts to gauge levels of metacognitive development in the engineering classrooms?

This research is part of a larger NSF-funded project that seeks to develop and pilot an intervention to help instructors gauge their students' metacognition levels in the context of engineering classrooms. Details of the parent study have been published elsewhere (e.g., Cunningham, Matusovich, Morelock, & Hunter, 2016; Cunningham, Matusovich, Hunter, & McCord, 2015; Cunningham, Williams, Matusovich, & Bhaduri, 2017)

2. Background

My research hinges on two propositions that are supported by extant literature. First, that engineering students' metacognitive development is important and that this metacognitive development may be hard to measure through use of surveys or interviews. Due to the challenges in measuring metacognitive development, it is recommended that instructors assess the development through analysis of student responses to reflective prompts. The second proposition relates to the challenges in efficient assessment of reflective prompts. As introduced earlier, although analyzing reflective prompts may be effective in understanding the student metacognitive development, for instructors to assess manually student responses may not be a time or resource efficient task. Deeper understanding of these two propositions will help us situate the need for a Natural Language Processing and Machine Learning based automated classification system to help instructors assess student responses, and thus facilitate metacognitive development in engineering classrooms.

2.1 Student Metacognition: Why is it important and how can we best measure it?

Metacognition is a construct that involves knowing about and regulating one's own cognitive processes. Cunningham et al. (2015) argue that metacognitive skills are significant for student learning and conceptual change, by enabling students to take ownership of their learning. It has been shown that students can benefit from developing

metacognitive skills inside the classrooms to become more skilled learners (Pintrich, 2002; Veenman et al., 2006). Becoming skillful learners is especially relevant for engineering classrooms as it may ultimately help students solve complex, ill-structured real world problems that are typical to the engineering practice (Cunningham et al., 2017).

Metacognition has been recognized as difficult to measure (Matusovich & McCord, 2012), and is typically studied directly (e.g., Veenman et al., 2006), or as a component of other frameworks such as self-directed learning (van Hout-Wolters, 2000) or self-regulated learning (Winne et al., 2006). Specifically, one of the challenges in studying/facilitating metacognition is that since the thinking happens internally for the student, it is very difficult to understand the metacognitive development through surveys or interviews. The challenges of studying metacognitive development through surveys or interviews may be attributed to the construct validity in the former and inability to have conscious access to discuss engagement in certain metacognitive practices in the latter (Baker & Cerro, 2000; Pintrich, Wolters, & Baxter, 2000).

Not only is metacognition tough to measure, there is also ongoing discussions in existing literature on *how* best to measure it. Direct and online assessments for metacognition, such as student self-reported think aloud protocols, as opposed to their offline counterparts such as questionnaires and oral interviews, may have higher preference among researchers (Veenman, 2005) because of the higher predictive capability of such assessments to indicate student performance since they often capture what students are thinking in the moment. However, Cunningham, et al. (2015) describe

how think aloud protocols and direct observations too may be disruptive to a students' learning process, since it often pulls students out of their normal learning environment. Cunningham, et al. (2015) thus build an argument supporting use of reflective written assignments from students to best learn about their metacognitive level, while also helping them develop the skill further. Consistent with Cunningham, et al. (2015), I argue for use of student reflective assignments to gauge their engagement in metacognition for an engineering class.

2.2 An overview of the increasing use of automated Text Analytics

For classroom settings, with enrollments ranging anywhere from 25 students to 300 students per instructor, analyzing responses to open ended assignments is difficult due to time and resource constraints, and often instructors are unable to give a timely and effective feedback to the students thus losing out on opportunities for developing their skills (Cuseo (2007); Soledad et al. (2017)). To address this problem, hiring additional faculty may not always be an option. Moreover, often for smaller classrooms even with a strength of 25-30 students, for a single instructor to manually grade or make meaning out of open-ended responses may be a time and resource intensive endeavor. Machine learning and Natural Language Processing, as demonstrated through this research, are a resource-effective alternative to help individual instructors, in both large and small classrooms, analyze textual data generated from student artifacts, and generate trends and

patterns, for example, to understand students' levels of engagement with the materials in the course.

Automated Textual Analytics systems such as the one proposed in this study are based largely on Natural Language Processing techniques, and often borrow from machine learning tools. Bird et al. (2009) attribute the term Natural Language Processing to cover any kind of computer-based manipulation of a natural language. Importantly, Bird, et, al. distinguish between natural languages such as English, German or Hindi from what they connote "artificial languages" such as those used in programming, and observe that the former have evolved over time, passing from generation to generation with rules that may be hard to pin down explicitly (pg.1).

Natural Language Processing may be used in conjunction with machine learning tools. While Natural Language Processing uses techniques to process language to a format recognizable by the machine thus making it easier for the machine to conduct statistical operations on it, Machine Learning techniques rely on statistical tools to enable the machine to iteratively learn from the data. In conjunction, these techniques have been used for varied tasks. For example, while Savoy (2015) used Natural Language Processing techniques to process word and phrases in a study to cluster 300 State of the Union Addresses to categorize presidential authorship; it has also been used by Pestian et al. (2010) to use the information from the Natural Language Processing step to then train a machine learning based system to classify genuine suicide notes from a corpus which included elicited ones. Pestian et al. (2010) found that their Machine Learning algorithm outperformed the mental health professionals by correctly classifying 78% of the times

(as opposed to 63% by the mental health professionals), whether or not a suicide note was genuine. These tools are more recently being widely accepted in educational interventions and research. For example, in understanding student vocabulary specific to a course (e.g., Variawa et al. (2013)), predicting student persistence in MOOCs (e.g., Robinson et al. (2016)), as a tool for student assessment (e.g., Magliano and Graesser (2012), or in making meaning of end of semester course feedback in large engineering classrooms (e.g., Soledad et al (2017))).

Thus, an example of a Machine Learning and Natural Language Processing related task is using a current data set to train an automated classifier. This is observed in the above example with the suicide notes classification. Typically, the task of classification involves learning from an existing labeled, commonly a human coded (i.e., labelled) data set, and then based on this learning automatically classifying a second data set of non-labeled data. Classifiers have been developed using machine learning and Natural Language Processing to classify smokers, i.e., smoker, current smoker, past smoker, non-smoker, and unknown based on text in the documents related to medical records (Savova, Ogren, Duffy, Buntrock, & Chute, 2008). For instance, in their study reporting on development of an automated Natural Language Processing based classification system for exam questions based on Bloom's Taxonomy, Jayakodi, Bandara, Perera, and Meedeniya (2016) indicate the accuracy of the system in correctly classifying the dataset to be over 70%. This percentage of accuracy indicates that of the number of times the automated system was used to classify the exam questions, it was successful in correctly classifying the questions seventy percent of the times.

In my study, I extend the scope of Machine Learning and Natural Language Processing based classification to enable classifying a new set of student responses into low, medium, or high based on expressed characteristics of metacognition. These labels generated by the automated system were then be compared against expert ratings of the same responses, to ascertain the level of accuracy of the classifier developed. For this study, I have predominantly used supervised Machine Learning algorithms combined with Natural Language Processing. Supervised Machine Learning refers to development of an intelligent system by providing it with labeled training data, thus enabling it to learn iteratively from pre-labeled data. Here, a training data set can be understood as a set of data that has been manually coded by the researcher (Scharkow, 2013). In the following sections, I will describe how previously coded and labeled responses were used to train a Machine Learning based classifier using Natural Language Processing features to predict the metacognitive level of a student written response to an open-ended prompt.

3. Method

Student responses collected as part of a metacognition development intervention served as data for this study. This intervention was developed to help instructors incorporate metacognitive modules within their courses to teach students about the critical elements of metacognition. Specifically, a six module intervention was designed for implementation in an engineering classroom (described further at: Cunningham et al., 2016; Cunningham et al., 2015; Cunningham et al., 2017).

An overview of the intervention will help understand the context of the study design. The intervention, as will be elaborated upon in subsequent paragraphs, comprised a total of six modules. Responses to the open-ended reflective questions embedded in modules one and three were used as part of this study. These responses were hand-coded by metacognition experts, and then used to develop the automatic classification system through steps detailed in the following sections.

3.1 Intervention Overview

The modules developed as part of the intervention address important metacognition concepts such as planning, evaluation, monitoring and control, through a video operationalizing elements of metacognition within an engineering context, followed by reflection questions, an in-class activity and a post class assignment (Cunningham et al., 2017). Table 1 provides and overview of the modules.

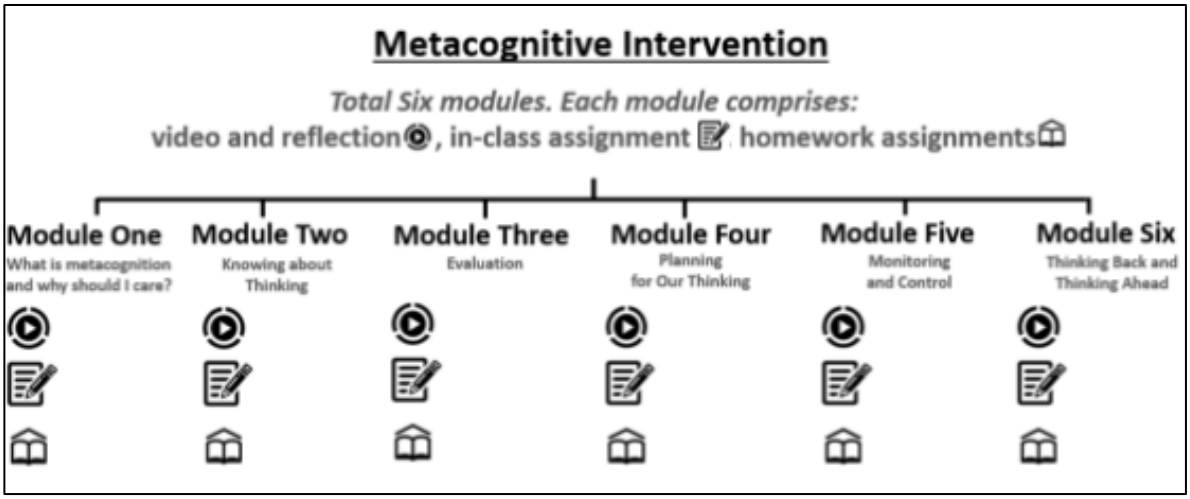
Manuscript Two. Table 1 Overview of the six modules in the metacognitive intervention

(Modules 1 and 3 included in this research)

Module	Title	Purpose
1	What is metacognition and why should I care?	Introduce students to the metacognition framework and argues for importance of metacognitive knowledge and regulation
2	Knowing about Thinking (Knowledge of Cognition)	Focus on metacognitive knowledge of self, tasks and strategies
3	Reflecting on Our Thinking (Evaluation)	Introduce students to the idea of assessing a learning experience to determine what worked and what did not
4	Planning for Our Thinking (Planning)	Introduces students to the idea of focusing on tasks that are part of big project and part of important gorals rather than tasks that are distractions

5	Optimizing Our Thinking (Monitoring & Control)	Introduces students to monitoring and controlling their learning during a learning experience, operationalized through Kolb's experiential learning cycle.
6	Thinking Back and Thinking Ahead	Serves as a summary that asks students to reflect on topics from the prior weeks and think about how they can apply what they have learned going forward

The goal of the overarching parent study (Cunningham et al., 2017) was to create metacognitive indicators based on the student responses that instructors could use to gauge individual and collective class progress in developing metacognitive skills. As can be seen from the Figure 1, each module includes a video operationalizing elements of metacognition within an Engineering Education context, reflection questions following the video, an in-class assignment, and a homework assignment. Students were asked to watch the video prior to attending class and submit responses to reflection questions. The in-class activities asked students to engage further with the content to build understanding. The follow-up homework assignments generally asked students to apply the content. To facilitate ease of use of the modules, they were designed to be short and customizable so as to fit with current content rather than forcing instructors eliminate course content to accommodate the modules.



Manuscript Two. Fig. 1 Overview of the metacognitive intervention.

(Modules 1 and 3 included in this research)

3.2 Site

The metacognition intervention was piloted in Fall 2015 at a small private engineering, math, and science school (PriI) in an engineering problem solving course. There were a total of eight sections of this course, with each section comprising 25-30 students, of which two sections received this intervention. The intervention was included as a graded element of the course, counting for a modest part of the homework grade. A version of the intervention was then repeated the following summer semester at a large public engineering college (PubU), where it was implemented in two introductory first year fundamentals of engineering course. The following Fall, the implementation was again repeated at PriI. For each iteration, the classrooms included engineering students being taught metacognition related concepts. The similarities in the iterations led to use

of the data from all three iterations collectively. Thus, while the first set of data was collected at a small private teaching focused institution, the second set of data was collected at a large public research focused institution. The two sites were intentionally diverse to promote transferability of the intervention across sites. The research team wanted to ensure that the metacognition intervention that was developed and piloted in one context was transferable to other engineering education classrooms in contexts which were different. Similarly, participants from different grade levels were also chosen to ensure that there was transferability of the intervention not only across sites, but also across grade levels for the same site. Table 2 below provides an overview of the sites and site characteristics.

Manuscript Two. Table 2 Overview and Characteristics for Sites

Semester	Site	Site Characteristics	Intervention Implemented For:	Number of Responses
Fall 2015	Pri I	small, private, engineering, math and science school	• Two sections of an engineering problem solving course. • At sophomore level. • 25-30 students per section. • Intervention was included as graded element of course.	73
Summer 2016	Pub U	College of Engineering part of a large, public university	• Two sections of a fundamentals of engineering course. • At freshmen level • 25-30 students per section • Intervention was included as graded element of course	41
Fall 2016	Pri I	small, private, engineering, math and science school	• Two sections of an engineering problem solving course. • At sophomore level. • 25-30 students per section. • Intervention was included as graded element of the course.	38
			Total Participants:	**152**

3.3 Participants

Questions from Modules One and Three were chosen for this analysis as they were deemed to be similar enough to analyze together. My dataset included 152 responses (Table 2) from students which were complete, and those which had associated student consent allowing use of the data for research. Data collection was approved by the Institutional Review Board (IRB) at both the sites, and for all implementations. Race, ethnicity and gender of the participants are intentionally not specified in this study so that participants are not potentially identifiable by the instructors who were not part of the research team. In order to maintain intentional anonymity of the responses, none of the descriptors for race or gender, which could then be used to trace back to the individual students, were recorded. Although exact quotes were used for the analysis, exact quotes are not presented in this manuscript to similarly protect participant anonymity.

The 152 collected responses were then hand-coded by researchers and assigned labels: high, medium, and low based on metacognitive level as deciphered from the response. Details on the responses from each of the three sites is presented in the Table 3 below.

Manuscript Two. Table 3 Overview of participants per module and site

(Total of 152 responses collected)

Site Overview		**Number of Responses**						
Semester	**Site**	**Module One (=76)**			**Module Three (=76)**			**Total (152) by each site**
		High	**Medium**	**Low**	**High**	**Medium**	**Low**	
Fall 2015	PriI	4	18	17	7	14	13	73
Summer 2016	PubU	1	12	7	4	9	5	41

Fall 2016	PriI	2	4	11	2	8	14	38
Total Per Module		7	34	35	13	31	32	

3.4 Data Collection

The data collected for this study comprised 152 student responses (Table 3) along with their corresponding labels (High, Medium, or Low) assigned to the responses by a team of metacognition researchers working individually on the coding. Coding was conducted using a priori codes Patton (2005), which were based on the metacognition framework from literature: "i.e. knowledge of persons, tasks, and strategies, and regulation of cognition". Cunningham, et al. (2017) describe the process of assigning labels for High Medium and Low metacognition for individual responses as:

> "the responses were ranked "high", "medium", and "low". A "high" response answered the question, described their strategies clearly and provided concrete, demonstrable evidence to their claims. A "medium" response has a clear strategy description and weak evidence to support their claims. A "low" response has vague description of strategies and no evidence to support their claims." (pg. 5)

Following coding of all responses by the primary researcher, another researcher on the team independently coded a sub-sample of the responses, and the labels (of High, Medium, and Low) for metacognition level assigned by both the coders was compared. The researchers discussed their individual ratings and compared the labels until consensus was reached on the labels generated for the responses. These labels generated

formed the dataset for the classification algorithm developed in my study. Examples of responses for both modules are presented in Table 4 below.

Manuscript Two. Table 4 Responses for Both Modules

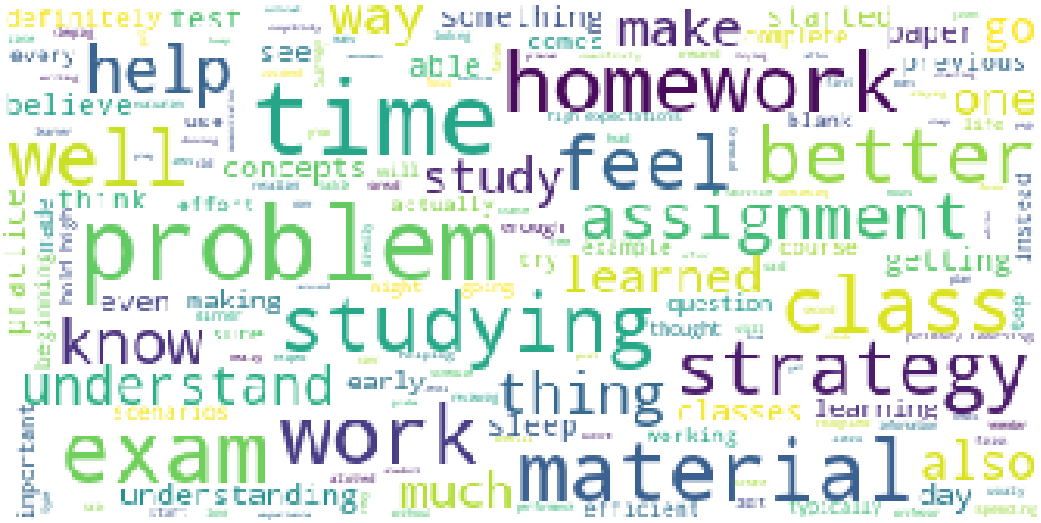
Module	Sample Prompt	Example of Responses to Prompt, Assigned Labels As		
		High	**Medium**	**Low**
Module One: **What is metacognition and why should I care?**	Please explain your selections from the previous question on the statements that most exemplify your learning experiences. For example, if you selected, "I hold high expectations for my performance in classes," explain why you hold these expectations.	'I have set high expectations from myself based on my high scores and good performance in the past. I know that if I am more organized and schedule time for regular study and cover a wide range of topics I will do well. I also set aside specific time for study. This helps me not miss other important things such as spending time with my friends and family, and for extracurricular activities. Putting aside time for studies, on a regular basis helps me feel like I am on the top of my class work. '	'I give my best effort. However, I do not set time limits when I am working on homework assignments. I keep working on the homework until it is done. Sometimes I have to stop my work if it is taking too long or if it is late. Usually I am able to finish my work early'	'I don't plan. Once I get back home from school I start working on my homework and finish it. If there is time left for other things I do them later.'
Module Three: **Reflecting on Our Thinking (Evaluation)**	Name one new thing you have been doing since completing the GAMES© survey in #2-Knowing About Thinking. How is it helping you be a more skilled and efficient learner? Review your plan and strategies to implement your plan from you submission to the previous assignment if necessary.	'In the past it has been more helpful for me to meet the professor when I am stuck on a problem for hours. Talking to the professor helps me eventually solve the problem. This strategy has allowed me to become more efficient in completing my homework and in helping me develop a better understanding of the problems themselves and the multiple ways in which they can be solved.'	'Recently I have started speaking more with friends about what we are learning in class. Discussing with them helps me understand what I may have missed. Another thing I have been trying is to recall concepts rather than just read through notes.'	'I prepared well for the exam, as a result of which I was able to think more efficiently while taking it.'

To further assess agreement between human coders, I requested an independent metacognition expert to code the entire dataset, and then compared the codes generated by the expert to the labels generated by the researchers. It was found that a 78%

agreement existed between the labels assigned by the researchers and those assigned by the expert in the two independent iterations. The goal of my study was to enable automated classification of the student responses into three levels based on metacognitive development, with an accuracy comparable to that of the human inter-rater agreement.

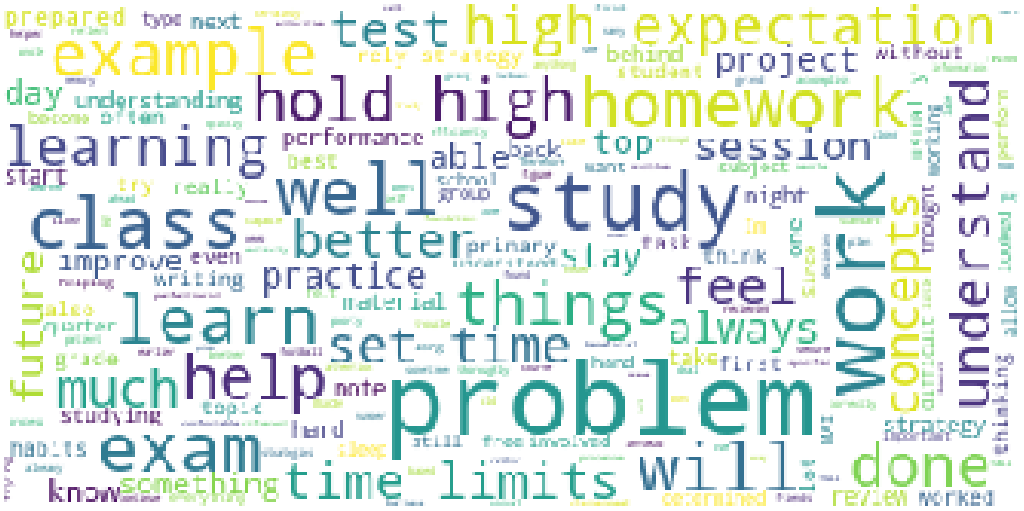
3.5 Data Analysis

The first step in my algorithm development was to seek to understand the data generated as part of the student responses. To help view word occurrences for the three levels of metacognition in the student responses, I generated word clouds. Weighted word clouds is a popular text visualization method that allows a quick overview of the predominant words in a corpus. I then used the information from the word clouds to develop features for my classification system. The word clouds for the high, medium, and low responses are shown in the Figure 2 a-c below.



2 a) Responses Scored as

2 b) Responses Scored as Medium



2 c) Responses scored as Low

Manuscript Two. Fig. 2 Word clouds for a) High, b) Medium, and c) Low Metacognition Response

From the word clouds, it can be seen that there are similarities and dissimilarities in the choice of words used in the three categories of responses. While words like problem, studying, and strategy show up as some of the most frequently used words for the high metacognition responses, for the medium metacognition responses both problems and studying seemed to be less frequently used. For the low metacognition category, the word problem again showed up as one of the most frequently used words, along with help, time, limits, better, expectation, etc. The word clouds thus help provide a sense for the word choices used in the three categories. However, visualizations of mere word frequencies may be misleading, since they only help visualize the words or phrases on the

basis of frequency of occurrence. In my next step to visualize the data, I went beyond word frequencies to look at the normalized frequency of occurrence using the tf-idf technique.

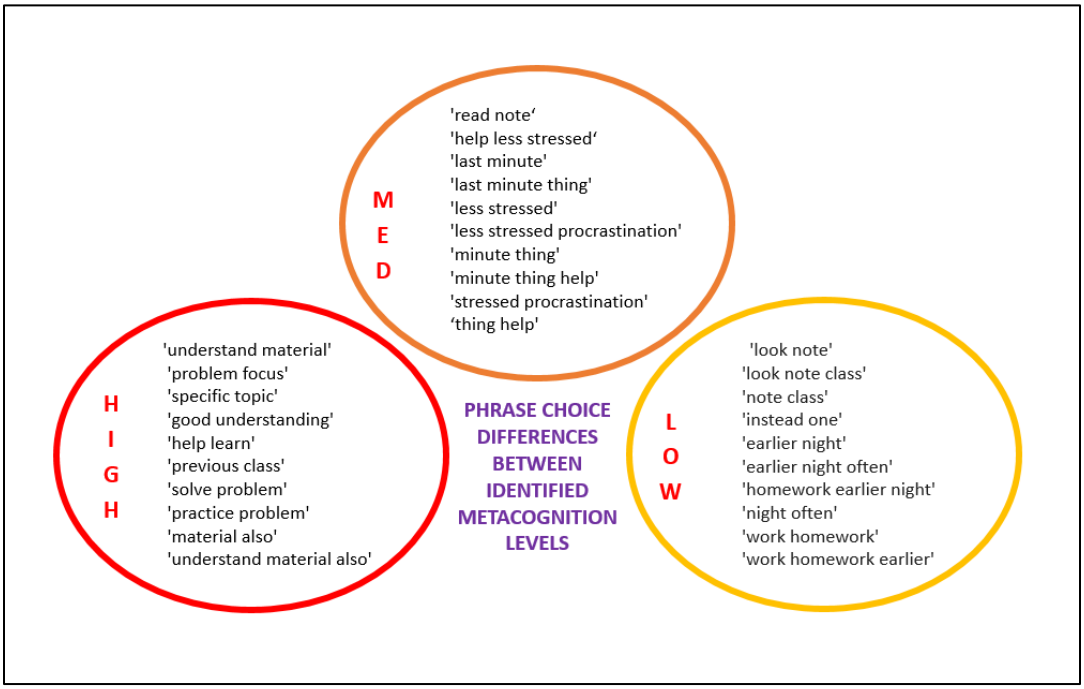
In analyzing large text corpora, tf-idf is a popular method which goes beyond the word frequency to not only compute frequency for the word in a particular document, but also multiplies this frequency by the inverse document frequency. Multiplying by inverse document frequency leads to lower tf-idf scores for words such as articles (a, an, the) which occur in large instances in both single documents, and in the entire corpora (Shi, Xu, & Yang, 2009; Variawa et al., 2013). Variawa (2014) described the computation of tf-idf scores:

> "The TF is a number determined by counting of occurrences of a particular word, and dividing that number by the total number of words in the target document: as such, it is a measure of frequency. The IDF is a measure of how important a particular term is within a set of documents, and is calculated by dividing the total number of documents by the number of documents in the set which contain that term, and then takes its logarithm. The TF-IDF formula multiplies these together and attaches the resulting score to each unique word in the target document." (pg. 204)

I used tf-idf to generate the top ten bigrams (set of two corresponding word tokens) for each of the three metacognitive level labels as assigned by the researchers. For module Three, which focused on the student responses to developing evaluative skills and reflect on their thinking, the tf-idf bigrams generated are shown in the Figure 3 below. It can be

seen that a majority of the high metacognitive responses shared instances of "understanding material", or "solving the problem", while phrase choices for the medium metacognitive level students indicate a procrastinator undertone with emphasis on "last minute" and "less stressed" emerging as top phrases.

This list of generated phrases was then shared with the team of experts, for an attempt at researcher triangulation, and to also determine if the subtle differences in word and

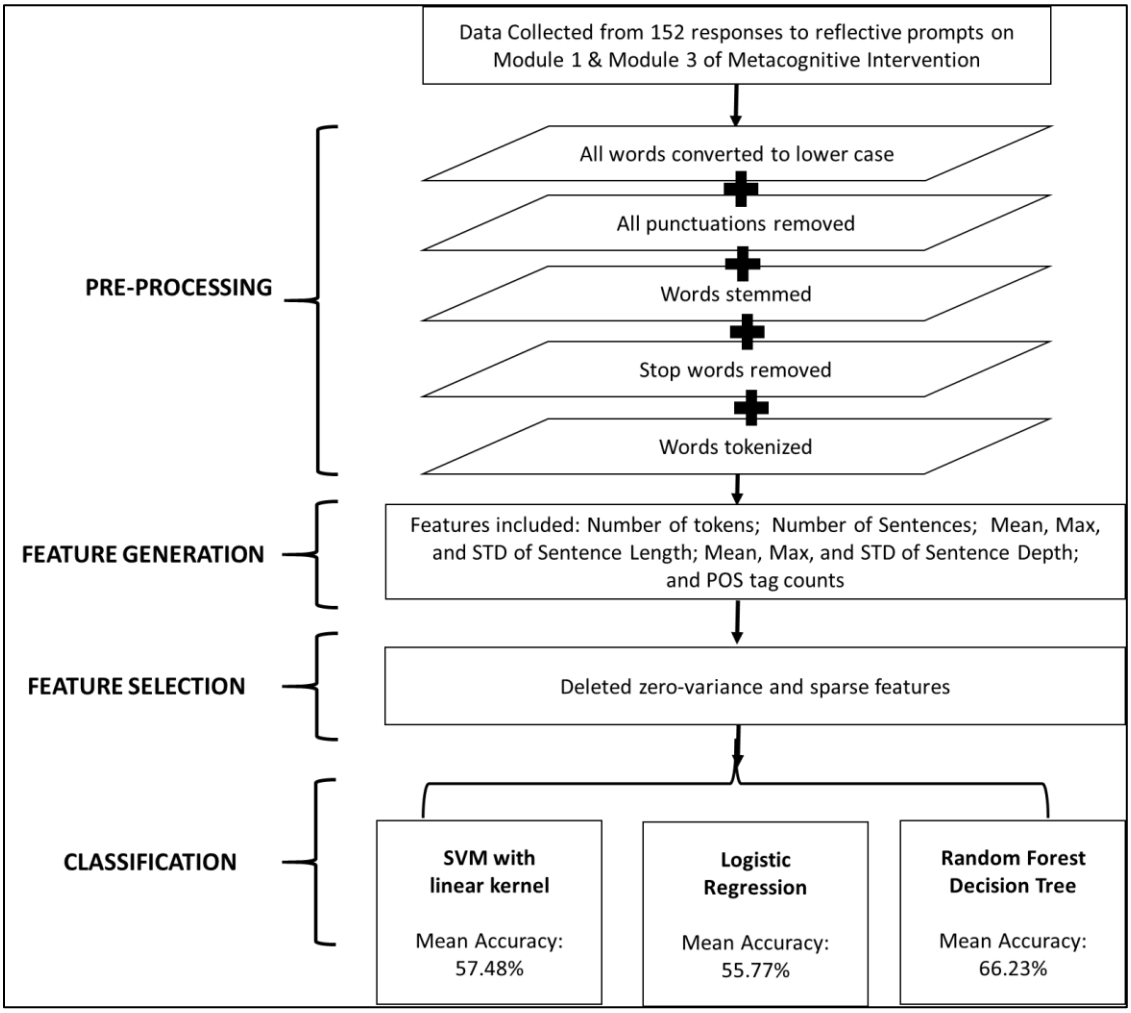


Manuscript Two. Fig. 3 Tf-idf bigrams highlighting differences in student responses for the three expert assigned levels for Module Three

phrase choices among the students from the three different groups made sense from a metacognitive standpoint, which was confirmed by the experts. Thus, I then proceeded with development of an automated classification system for the complete dataset.

3.6 Automated Classifier Development

The 152 student responses, from data collected for three iterations of the intervention, along with the corresponding expert assigned level indicating metacognitive development for each response collected were then used for developing and testing the automated classifier. NLTK (Natural Language Toolkit) which is a suite of open source Python modules (citation) was used for the classifier development and testing. Figure 4 below shows the different steps in algorithm development and testing.



Manuscript Two. Fig. 4 Steps in Classifier Development

The following paragraphs delve deeper into the description for each stage of the process.

Pre-Processing. The first step in the algorithm development was the pre-processing, which included preparing the data for classification. Pre-processing steps are common to machine learning and Natural Language processing research and typically involve removal of punctuation marks, converting to lower case, and then reducing words to their lemma or stem word (e.g., Soledad et al. (2017), Bhaduri & Roy (2017)). At first, all of the punctuation marks were removed from the document and all the words were

converted to the lower case. I then used a stemmer to further stem the words to their root. A document can contain different forms of the same word (e.g. organize, organizing. organizes) which for analysis purpose should be considered as a single token. For example, after stemming, all the words in the above example will reduce to “organize”. Many words in the document (e.g., articles such as the, a, an) often provide no additional information about the content. These are termed as stop-words. Thus, the final step in pre-processing was to use the in-built NLTK feature to remove English stop words and tokenize each word.

Feature Generation. Features can essentially be understood as the basis for classification of the responses by the algorithm. They can also be understood as an array of unique numbers that can be used to define a response, and distinguish it from the next. For my classification system, I used five types of features to build an array of length 43 per response. These features were developed based on literature (e.g., Pestian, et al. (2010), Savoy (2005)). The types of features used were:

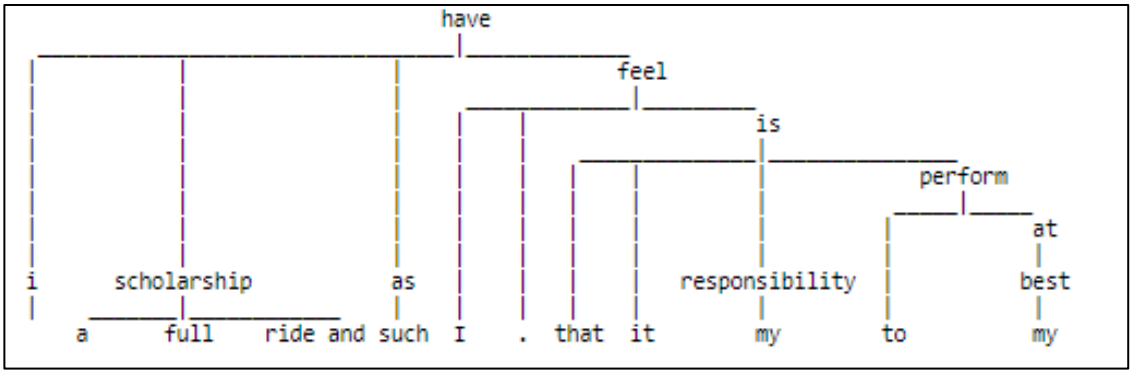
a) Token Numbers. Similar to the use by Pestian et al. (2010), the number of words (unigrams) that remained in a response after the stop words had been removed was considered as token numbers. This was normalized per response, by dividing by the total number of original words in the response.

    Thus, if a response were to comprise a single sentence: “I have a full ride scholarship, and as such I feel that it is my responsibility to perform at my best”. The number of normalized tokens for this sentence would be 13/21 = 0.61.

b) Sentence Number. The sentence number type of features used information about the number of sentences per response. It also calculated a mean value for number of sentence for all the responses, and then computed the standard deviation for each response.

c) Sentence Length. The sentence length type of features used information about the number of words per sentence per response. I also calculated a mean value for length based on number of words of the sentences for a response, and then computed the standard deviation for the sentence lengths for the given response.

d) Sentence Depth. The sentence depth was established based on the dependency parser tree created for each sentence in the response. Dependency parsers present the relationship between words in a sentence (Duda & Stark (2001)). For a given sentence, “I have a full ride scholarship, and as such I feel that it is my responsibility to perform at my best”, Figure 5 shows the parsed tree structure. The depth of this sentence can be calculated as 7, which is the distance from the top node to the lowest node. Similar to other features, I calculated a mean, maximum, and standard deviation for tree depth for each response.



Manuscript Two. Fig. 5 Depth of a Tree

e) Parts of Speech. In addition to the above, 35 more features were added through in-built NLTK Parts-of-Speech tags generated at the word level for each sentence. For each of the responses, the counts for the parts of speech tokens were generated and divided by the total number of tokens to normalize the count.

Thus, using the five types of features described in this section, I had an initial count of 43 features for my classification task. The Table 5 below tabulates the initial features generated for the classification tasks.

Manuscript Two. Table 5 4 Features Used for Classification (9 features deleted from Parts_of_Speech)

	Feature Type	Description	Number of Features
1	**Tokens_Number**	The number of tokens (i.e., unigram words after removal of stop words from each response) normalized per response by diving by number of total words.	1
2	**Sentences_Number**	The number of sentences in the response.	1
3	**Sentence_Length**	The mean, standard deviation from mean, and maximum length of sentence per response.	3
4	**Sentence_Depth**	The mean, standard deviation from mean, and maximum sentence depth of each response.	3

5	**Parts_of_Speech**	Parts of Speech counts for the response.	35
	Total		43

Feature Selection. Once the initial features were set up, it was necessary to get rid of features that were not contributing to the classification tasks. I first eliminated 5 features which were zero variance, i.e., their values were unique throughout all responses, thus not contributing as a discriminatory feature. Next, I deleted 4 sparse features which, which only showed up in 10% or less of the dataset, and was zero value for the other responses. Deleting 9 features left a final feature count of 34. I got a matrix of dimension 152 x 34. Here, 152 was the number of responses and their corresponding labels or data points; and 34 the number of features. This was input into the classification algorithms for the task of classifying the responses.

4. Results

The 34 features were then input into three classifiers. In this section, I present a comparison of the results for the three classifiers that were used. Based on the results of this comparison, the classifier that outperformed the others was implemented into the final system. The performance for predictions in the three class and two class classification tasks are described, drawing from the student responses and comparing to the human researcher assigned labels.

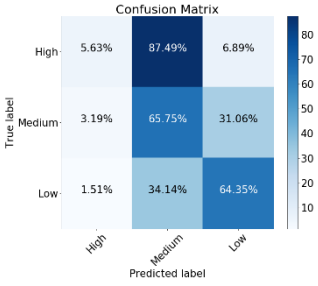
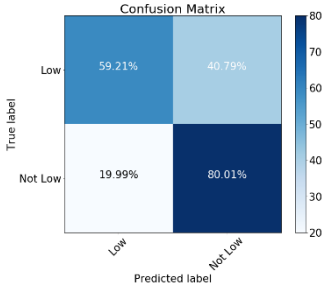
4.1 Comparing Results across Classification Algorithms

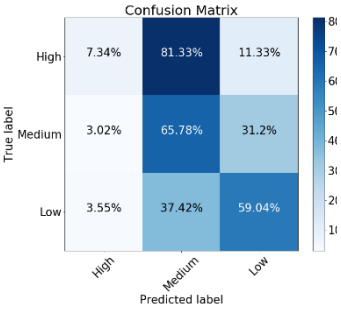
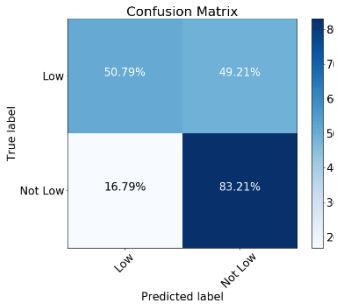
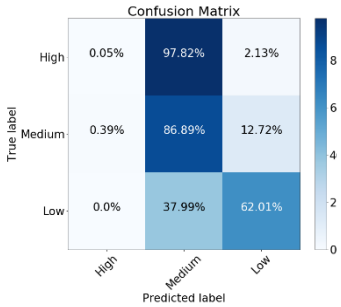
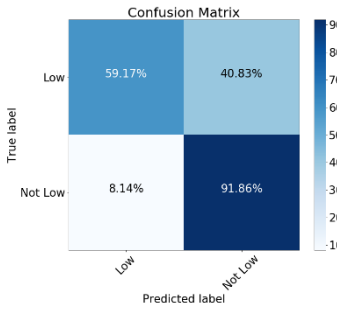
Three main methods were compared for the automated classifier. I trained and tested the dataset for classification using (1) Support Vector Machine with Linear Kernel, (2) Logistic Regression, and (3) Random Forest Method. All three of these algorithms are in-built functions in the NLTK library, and were used for the same set of features, for comparison. The complete dataset was partitioned into 80%-20% training-testing sets, such that for any given iteration of the algorithm, only 80% of the dataset and corresponding labels was used to train the classifier while the remaining was used to test the blinded results of trained classifier against the existing labels. 1000 iterations were conducted, and accuracy means over all iterations for each method noted. Table 5 below provides an overview of each of the three classification system, and corresponding confusion matrices to provide details about the performance for the dataset. The numbers in the confusion matrices were generated as means from 1000 independent training and testing iterations.

Two types of confusion matrices are presented. Confusion matrices can be interpreted through the elements in the diagonal. The positive diagonal elements from the element in the first row first column to the last row last column, represent the accuracies of the instances wherein the machine correctly predicted a class, as compared to the true value of the class. I ran the classification algorithm to classify student responses into high, medium, and low; as well as low versus not low. The latter was to help instructors identify students who may have submitted responses which are indicative of low metacognition level, so as to enable instructors to modify their instruction accordingly to provide more directed scaffolding in the metacognitive development process. As can be

seen the Random Forest method seemed to outperform the rest, with a three-class classification mean accuracy of about 66% averaged over 1000 independent runs.

Manuscript Two. Table 6 Performance of Classifiers

Classifier	Description	Mean Accuracy	Max Accuracy	Confusion Matrix of Performance (Tri-class and Bi-class)
SVM with linear kernel	Support Vector Machines (SVM) were developed as a supervised machine learning classification technique by Vapnik and co-workers (Vapnik, 1995). They separate a nunary labeled training data with a hyperplane that is maximally distant from them (Furey, et al. (2000);Joachims (1998)). A linear kernel was used for the text classification due to the large number of features.	57.48	80.64	 

Logistic Regression	Logistic regression classifier is a discriminative classifier which works by extracting weighted features from an input, combining them linearly, and using this information to apply a function to classify text (Jurafsky & Martin, 2016).	55.77	83.87	 
Random Forest	Developed by Breiman (2001), random forest classifiers grow branches of decision trees based on subspace of features from the training set, for each node.	66.23	87.09	 

4.2 Discussion

As described in an earlier section, I used my classifier to automate classification for a tri-class classification (i.e., classifying as high, medium or low for metacognitive development), and bi-class classification (i.e., low or not low for metacognitive development). In this section I explain the results of the classification system presented earlier by presenting a discussion drawing from the responses that were classified. I will then explain the limitations of this work in terms of the challenges in working with small datasets. Finally, I will present how a practical implementation of this classification can be used by an instructor, keeping the instructor in the loop of the classification process by taking their inputs on whether or not a response seems appropriately classified.

4.2.1 Tri-class classification problem

The tri-class classification problem can be understood as using an algorithm to classify the text responses into either High, Medium, or Low. In this section, I will discuss the results from the Random Forest classifier, since it outperformed the other classifiers. As can be seen from Figure 6 below, the Random Forest classifier was successfully able to classify with 86.6% accuracy the medium responses as medium. An example of such a classification is the response:

> "I was well prepared for the first exam. I did study, but took time on my homework assignments to make sure I first understood the concepts that were involved. On the night before the exam, I used the page of notes as a chance to summarize the relevant concepts and make connections between them. I was able to then gauge my proficiency for the material mainly

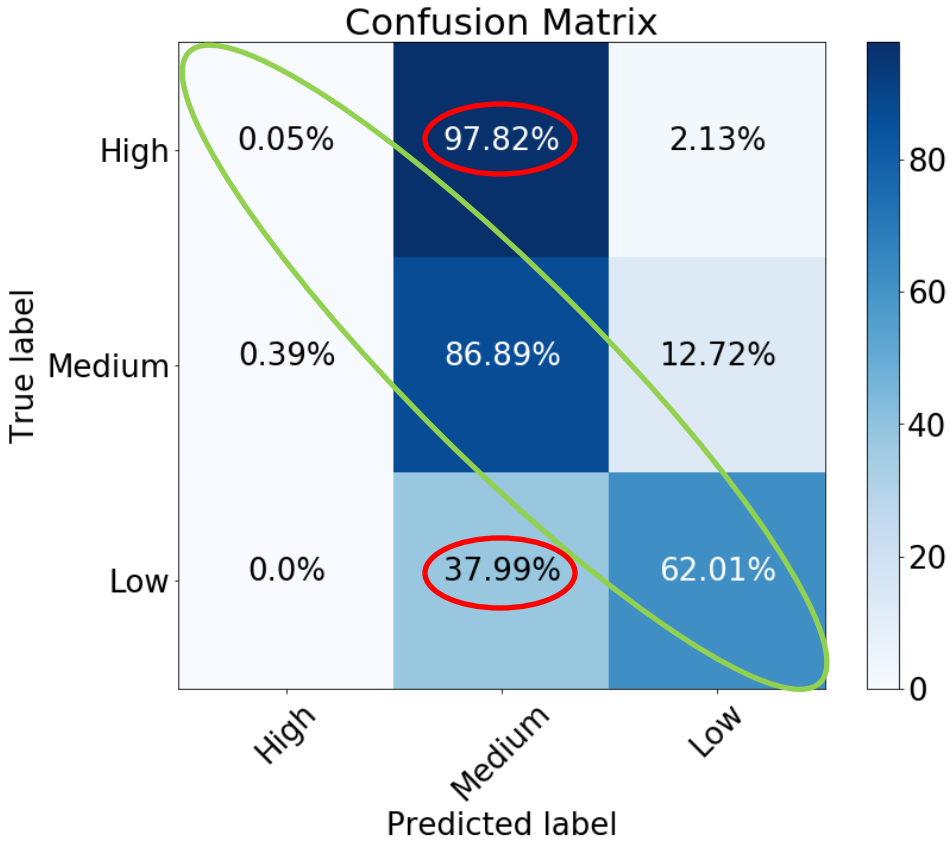
> through my grades on homework problems. These problems seem to be a good way to point out my more major weaknesses, but may not be comprehensive."

The classifier was successfully able to classify this response as a Medium in terms of metacognition level, which matched the label assigned to this response by the human researcher. However, there were few responses which were labeled as Medium by the human researchers, but which were falsely classified by the machine as either High or Low. The mean accuracy for falsely classifying the responses labeled as Medium as High was very low at 0.3%, while the instances for misclassifying them as Low was 13%. Consider as an example the response:

> "My primary learning strategy for most courses is to solve the problems on my own, without looking at similar examples. That is, unless I get very stuck. Often, I reflect on my problem solving strategy and how it worked out. I choose to rely on this technique because I think that by struggling through a problem and trying techniques that may not work out I learn more about why one thing might work and another might not. This is better rather than just seeing a situation and thinking "This situation always calls for this technique". I also avoid working with other people mainly because I want to make sure I get to do all the work in solving the problems. I rely on this strategy a lot because it has worked well for me."

This response was classified by my algorithm as High, however, the label assigned to this response by human researchers was that of Medium. This indicates that the machine may have found the sentence structure and length, the sentence complexities (as indicated by the tree depth), and part of speech counts for this particular response similar to those in the high category, as opposed to the medium. Figure 6 details the accuracies and misclassification percentages for other combinations. Overall, my system was able to accurately match the human

rater assigned labels for the tri-class classification with a mean accuracy of 66% averaged over 1000 independent iterations, using the random forest classifier.



Manuscript Two. Fig. 6 Confusion Matrix for Output of Random Forest Classifier for three way classification (averaged over 1000 independent iterations).

In Fig. 6 above, if we follow the elements across the positive diagonal (encircled in the figure), we see that the algorithm was correctly able to classify 86.89% of the mediums, but only 62.01% of the lows and 0.05% of the highs (averaged over 1000 independent iterations). This indicates that the machine had a higher probability of correctly classifying a medium response accurately, as compared to the high and the lows. For both the lows and the highs, the machine mis-classified a majority of the responses falsely as

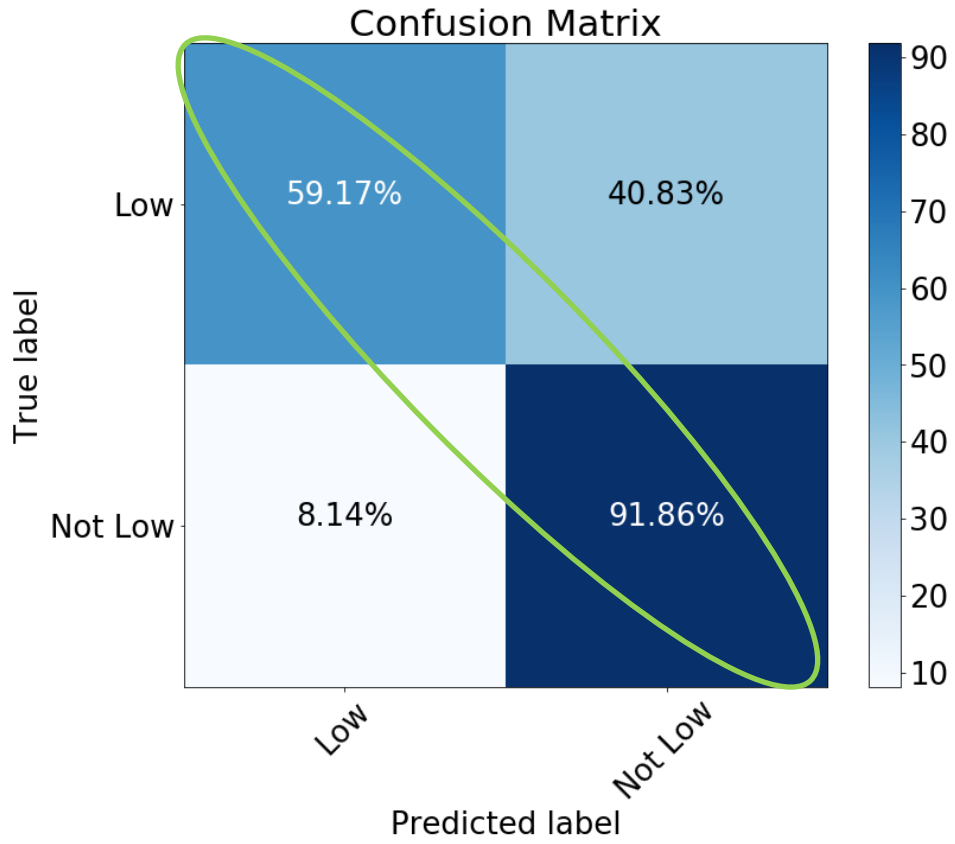
mediums (encircled in red, for ground truth based on human assigned labels). One reason for this misclassification may be the unbalanced training dataset, which is skewed in terms of frequency towards instances of mediums and lows, rather than highs.I present a small discussion in the limitations section on the training sample size dependence of automated classification system, and describe how and why the number of training instances may have impacted these testing accuracies.

4.2.2 Bi-class classification system

My automated system was developed keeping in mind instructors in large engineering classrooms. Often, instructors may want to get a snapshot of the class performance, to establish an idea of the percentage of the class that may be lagging behind, or may need additional scaffolding. In order to facilitate this, I trained my algorithm to perform a binary class classification to classify responses as Low, and Not Low. The purpose of this exercise was to isolate responses that may be low, so that the instructor may be able to direct their instruction to students who may be identified based on these low labels. An example of a response that the algorithm classified as low is:

> "I tried to understand how prepared I was based on my understanding of the homework and my recollection of certain concepts."

This response was also categorized as Low by the researchers, and thus there was an agreement between the ratings given by human and machine. Figure 7 details the accuracies and misclassification percentages for other combinations.



Manuscript Two. Fig. 7 Confusion Matrix for Output of Random Forest Classifier for two way classification

Again, in Fig.7 if we focus on the positive diagonal (highlighted in green), the percentages represent the accuracies for classification of the classes correctly by the machine. Thus, the machine correctly classified 59.17% of the low responses as low, and 91.86% of the not low responses as not low. These percentages are averaged over 1000 independent training and testing iterations of the classification. Overall, my system was able to accurately match the human rater assigned labels for the bi-class classification with a mean accuracy of 79% averaged over 1000 independent iterations, using the random forest classifier.

4.3 Limitations and Future Work

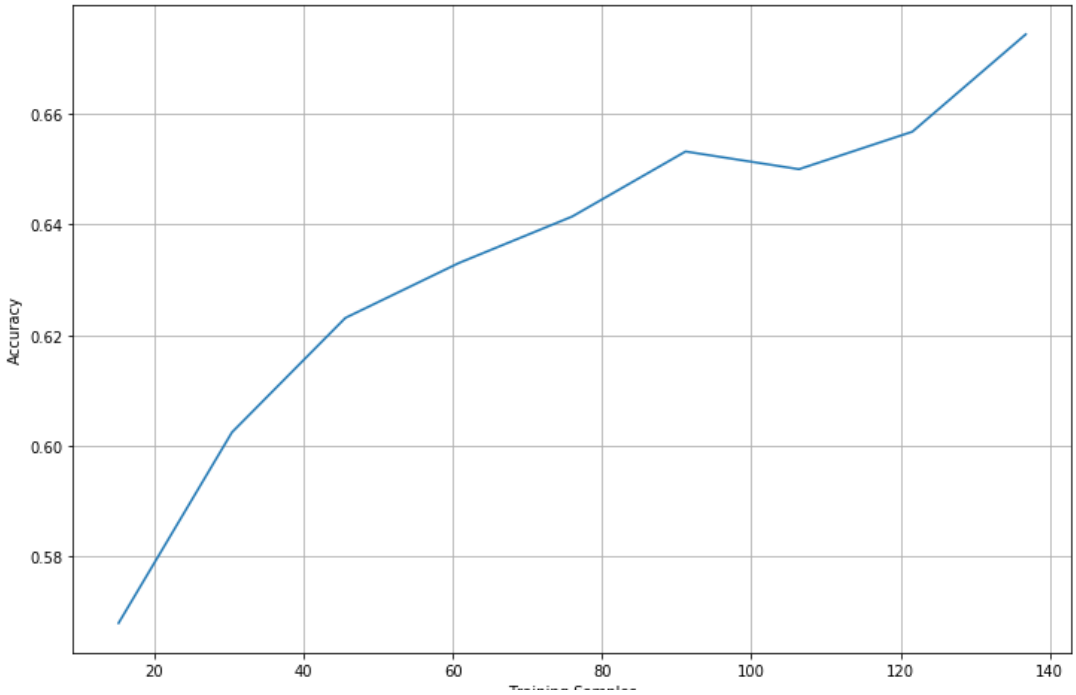
This work is limited by the size of the dataset and the generalization of results by the machine. One of the major limitations of intelligent systems, such as the classifier developed as part of my research, is their heavy reliance on the training data. Xu, Lan, Lu, Niu, and Tan (2012) in their paper on prediction of connectives using machine learning comment on the limitations of language models. They note that the quality of a predictive language model, in terms of the quality of the results, depends largely on the training data used. Scharkow (2013) observes that supervised machine learning methods are unsuited for rare categories because it takes a few dozen or even hundreds of positive training points to establish a stable statistical model for classification tasks. For this particular dataset, I observed an inadequacy of data points. This was especially true for the instances of High responses as compared to the total responses. Since the machine was not able to see adequate instances of high in the training set, it was more likely to mis-classify those instances as low or medium in the classification step. The inadequacy of data was further established when I found that increasing the training set (for example, from 75% to 80%) increased the overall mean accuracy. Figure 8 below plots the increase in the mean accuracy for the random forest classifier over 1000 independent training and

testing trial runs. Thus, a future step would be to increase instances in the training sample so as to better classify the test/blind responses.

Duda, Hart, & Stork (2012) explain another related limitation of automatic systems (such as classifiers) due to the central aim of generalization. They note:

> "The central aim of designing a classifier is to suggest actions when presented with novel patterns, i.e., patterns (or instances) not yet seen during training. This is the issue of generalization. It is unlikely that the complex decision boundary would provide good generalization, since it seems to be "tuned" to the particular training samples, rather than some underlying characteristics or true model." (pg.5)

In instances of inadequate training data, the system is not able to recognize an instance in the test data set, since it has not been observed in the training. In such situations the chances of the results derived from the system being wrong increases, thus lowering the quality of the results.

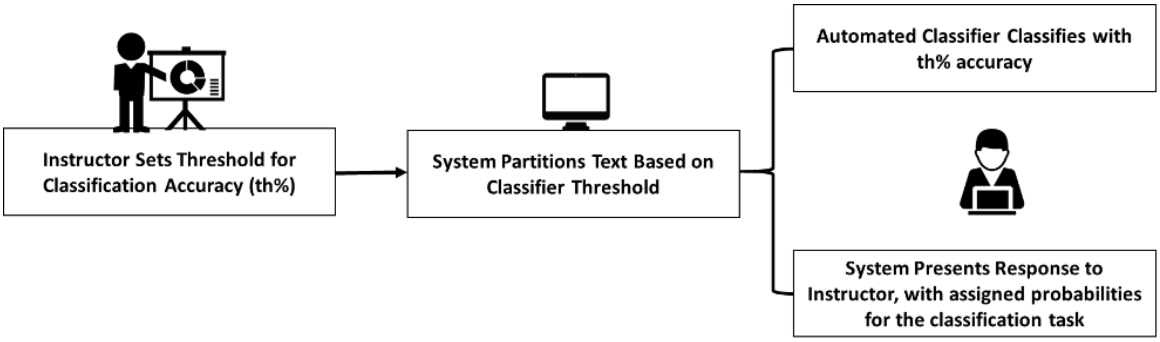


Manuscript Two. Fig. 8 Effect of Training Size on Mean Accuracy

For example, in this dataset, the algorithm may have defaulted to medium in instances wherein it could not discriminate whether a response was high. Keeping these limitations in mind, further research could continue to exploit the resource and time effectiveness of automated techniques using machine learning and natural language processing tools in the engineering classrooms by using automated analysis to aid instructors in facilitation and assessment of students' skill development in engineering classrooms.

4.4 Practical Implementation with Instructor in Loop

The challenges with establishing higher accuracies with smaller datasets such as for this project, as presented above may be alleviated through the use of feedback from the instructor in the classification process. This is an example of how my classification system can be used by the instructors in the classroom to semi-automate the classification process. For example, consider that the instructor wants to set a threshold for the machine, 'th%'. This threshold percentage represents the level of confidence the machine has in classifying a given response. Let us set this percentage at 60% for instance. Thus, the machine will report the classification results for all the predictions for which the confidence of prediction by the machine is 60% or higher. The rest of the responses, it will present to the instructor, with probabilities for classification for the responses, although the final decision for the classification for this subset of data will be with the instructor. Figure 9 below shows the flow of such a semi-automatic system.



Manuscript Two. Fig. 9 Semi-Automated Instructor in Loop System

For a threshold of 60 percent confidence prediction, my random forest classifier based system was able to classify 50.8 % of the test data set with an average time for decision at 0.014 second. The average accuracy for predictions for this reduced dataset was 73.19%. For this system, as the threshold of confidence for prediction is increased, the accuracy of the machine generated code will increase as well. A trade-off in this case is the number of instances that will be automated, which will decrease with an increase in the value of the threshold. An instructor can decide the level of accuracy they would like the machine to maintain and automate generation of the levels for a reduced dataset. The rest of the dataset can be coded by the instructor themselves. This instructor in loop system is presented as a proof of concept. Similar semi-automated systems with higher accuracies may be developed which can help the instructor automate coding for more than half of their response dataset, which will still be time and resource effective as compared to manually assessing all of the individual student responses.

4.5 Summary

Metacognition has been found to be important to students, helping them become more skilled learners (Pintrich, 2002; Veenman et al., 2006). The importance of metacognitive development is particularly true in helping students problem solve complex, ill-structured, real world problems which researchers have found to be typical to engineering practice. To help instructors in engineering classrooms to gauge the metacognitive levels of students through analysis of reflective assignments, as well as to provide feedback on such reflective assignments, it was necessitated that a time and resource effective method be implemented. Through this research I sought to answer the research question: "How can Natural Language Processing be used to automate the process of classifying student responses to reflective prompts to gauge levels of metacognitive development in the engineering classrooms, thus aiding engineering instructors in facilitating metacognitive practice/development?"

In this manuscript I have described how successful implementations of Natural Language Processing and Machine Learning based classification algorithms can prove to be a powerful tool to help analyze large volumes of textual data, especially in the engineering classroom. My research primarily focused on supervised machine learning based classification of student responses using Natural Language Processing based tools, providing an automated classification system which instructors may use in their

engineering classrooms to gauge, and hence better facilitate student metacognitive development. This research endeavor is particularly timely for engineering education contexts given the ongoing conversation on the challenges for instructors in present day universities to sustain the practice of giving timely and meaningful feedback to student cohorts, due to the increasingly large undergraduate student populations (Cuseo, 2007; Calvo & Ellis, 2010).

In my research I have demonstrated how machine learning and natural language processing were used in conjunction to classify student responses to reflective prompts with the best performance with a random forest classifier with tri-class mean accuracy of 66% averaged over 1000 iterations. The maximum accuracy for this classifier was found to be 87%, over 1000 trials. These accuracies may be found to be comparable to the human inter-rater agreement of 78% that was generated through manually coding and analyzing the responses. At the minimum, automated systems like the one described in my research, can help instructors gauge the classroom environment. The classification accuracy was found to increase to 79% for a bi-class classification for Low and Not Low (i.e., Highs and Mediums). Identifying students whose responses may be binned into the "Low" category for metacognitive development may be useful for the instructor to modify their instructional techniques. The instructor could then possibly give more directed instructions pointing students to resources which would be most beneficial for them to develop their skills.

4.6 Implications for Research

Through this study, I have demonstrated the use of a machine learning and Natural Language Processing based automated classification system to automate assignment of labels related to metacognitive engagement to individual engineering students' reflective assignments. Research such as this are important for researchers working on similar datasets to know what kind of techniques are suitable and have provided desirable outcomes. My choice of random forest as the classifier and the fact that it outperforming the other two classifiers, for example, can educate researchers working on similar datasets with the objective of automating the classification process. In the literature, there are several classification algorithms that are available to researchers. In text classification problems for example, SVM classifiers are popularly used (Chen, 2011). Chen explains how although SVM classifiers work well for high dimensional problems such as those afforded by text data, these are often memory intensive and hard to interpret. Chen advocates for ensemble methods such as random forest classifiers, which are easier to interpret, scalable and can easily handle interactions across features, which may be non-parametric. However, there is no "one size fits all" in cases of algorithm use, and, as Chen advises, researchers must try out several different classifiers ultimately selecting the one that outperforms the rest.

In prior work, it has been found that some researchers find making informed decisions from a wide range of options, in terms of techniques and methods, as one of the challenges of using Natural Language Processing for education research. This is especially true in instances wherein there is no researcher on large educational research

teams with expertise in algorithm development or use. In such cases, if the team wants to incorporate Natural Language Processing based techniques they may rely heavily on prior work in similar contexts. In such cases, providing results of use of Natural Language Processing techniques with adequate methodological transparency may be particularly useful to researchers, and ultimately for advancement of the field. Results of research studies such as this will help other researchers discern suitable methods as a starting point from which they can further build their work.

4.6 Implications for Practice

Using reflective metacognitive assignments in the classroom helps students better understand concepts and develop learning related skills. Automated systems for analyzing text-based assignments can be particularly beneficial to engineering instructors in large and small classrooms alike. In large classrooms, with the number of responses making it time-consuming for the instructor to assess, such automated systems can make the assessment process time and resource efficient. However, this system is not only relevant in large classrooms alone. The primary benefit of these systems is to provide individual instructors with increased autonomy to include more of open-ended reflective prompts in their instruction, without having to assess those manually. Instructors can make use of these automated assessment tools to engage students by assigning more reflective, open-ended prompts, and assessing those prompts swiftly to give the students more directed feedback. Although this particular study focusses on reviewing open-ended

responses to assignments related to metacognition, successful results in this endeavor may be helpful to pave the path for instructors to implement similar systems for assessing other skills as well.

Acknowledgement

This study is part of a parent study funded by the National Science Foundation under Grant Nos. 1433757, 1433645, & 1150384.

MANUSCRIPT 3: NLP IN ENGE RESEARCH

USING NLP IN NOVICE-LED EXPLORATORY QUALITATIVE DATA ANALYSIS

1. Introduction

Analyzing qualitative datasets manually is a time-consuming process, which the use of Natural Language Processing techniques can mitigate by making the process more time and resource efficient. Creswell (2014) elaborates on how qualitative data analysis involving hand-coding is a "laborious and time-consuming process even for data from a few individuals" (pg. 195). Teddlie and Tashakkori (2012) note that for larger sample sizes and more complex analyses computers may be more essential. Creswell (2014) and Guest (2012) describe the utility of computer based qualitative data analysis to organize, sort, and search for information in the text thus helping researchers to access the entire text more conveniently. In this manuscript, I introduce Natural Language Processing techniques as exploratory tools to analyze qualitative datasets, thereby helping researchers make meaning out of large volumes of interview transcripts.

Ease of organizing, sorting and searching for information in text, as available through the use of Natural Language Processing techniques, may especially be useful for researchers working on exploratory analysis of large textual datasets thus helping them

cluster topics, find word and phrase frequencies, and model the text. This is especially true for researcher who are novices with regard to topic and may not have a deep understanding of the topic of concern. At the same time, the role of novice researchers can be important as demonstrated through a paired novice-led thematic analysis approach (Montfort, Herman, Brown, Matusovich, & Streveler, 2013). In fact, Montfort, et al. (2013) claim that researchers working together to analyze qualitative datasets provide a richer and more rigorous analysis which is theoretically sound when the analysis is led or guided by a relative novice in the content area. My research expands upon the idea of novice-led paired thematic analysis which was introduced in the work by Montfort et al. (2013).

The purpose of this manuscript is to report on how insights from Natural Language Processing techniques were used to facilitate exploratory analysis of a large interview dataset by a novice researcher, which in turn contributed to a team of experts' understanding of the dataset, as well as presenting the expert researchers with directions for detailed and more nuanced in-depth qualitative analysis. Thus, for my research, I implemented the novice-led analysis on a dataset comprising interview transcripts of undergraduate engineering students talking about career preparedness. This dataset was developed as part of an NSF funded longitudinal study of early career preparedness and decision making among engineering students called the Professional Engineering Pathways (PEPS) study. The PEPS study was proposed by Brunhaver et al. (2015) to support an identified need for diverse and qualified engineers in the workforce by considering the choices made by recent graduates.

The specific research question driving this study was:

RQ: What insights can be gained from using Natural Language Processing techniques for exploratory novice-led analysis to inform opportunities for future analysis for a team of researchers working on understanding engineering student choices related to career preparedness?

Answering this research question helped provide expert researchers on the PEPS team with insights to direct further qualitative study of the students' career decisions. These insights were in the form of word clusters, word frequencies and topic modeling, which provided an overview of the qualitative dataset collected from interviewing students about their early career decisions. The insights from the analysis of the interview excerpts helped us understand how engineering students describe their career preparedness and early career decisions. These findings informed my meetings with the larger PEPS team. The results of this analysis led to newer directions for analysis for the PEPS team in their study of the early career decisions of the undergraduate engineering students. Studying these early career decisions, can be helpful to understand why graduates may decide to leave the discipline upon graduation despite having earned a degree in engineering. An understanding of choices made by engineering graduates may thus help the Engineering Education discipline work towards more directed efforts at retention of engineers in the workforce.

2. Background

To help motivate this research, in this section I l describe how Natural Language Processing is currently used for automated text analytics. An overview of the capabilities of automated text analytics using Natural Language Processing techniques will help establish how novices may use these techniques to facilitate their analysis of qualitative datasets. After introducing Natural Language Processing capabilities, I present a brief discussion on how novices may approach qualitative datasets, and how Natural Language Processing can help novices by presenting an overview of the dataset Finally, I will explain how both Natural Language Processing and novice-led exploratory analysis were useful for the PEPS dataset

2.1 An overview of Automated Text Analytics

Automated text analytics are becoming increasingly popular as means to understand and extract information from large volumes of text-based data. One of the reasons for this growing popularity for machine-based analysis of data may be attributed to increasing computational capabilities of machines (Baker & Inventado, (2014)), as well as the readily available techniques for digitization of data (Torii, Tilak, Doan, Zisook, and Fan, 2016). Digitization of data allows ease of data collection for analysis. Since the data is already in digitized format, the researcher can use it as an input for an automated system to analyze. Elaborating on the present era of digitization, Torii, et al. (2016) explain how analysis of such readily available, digitized, and large text-based data sets can help

researchers gain deeper insights related to certain contexts. Torii et al. (2016) used a machine learning approach to extract information from 1.3 million product reviews available online, to gain deeper understanding of health related issues.

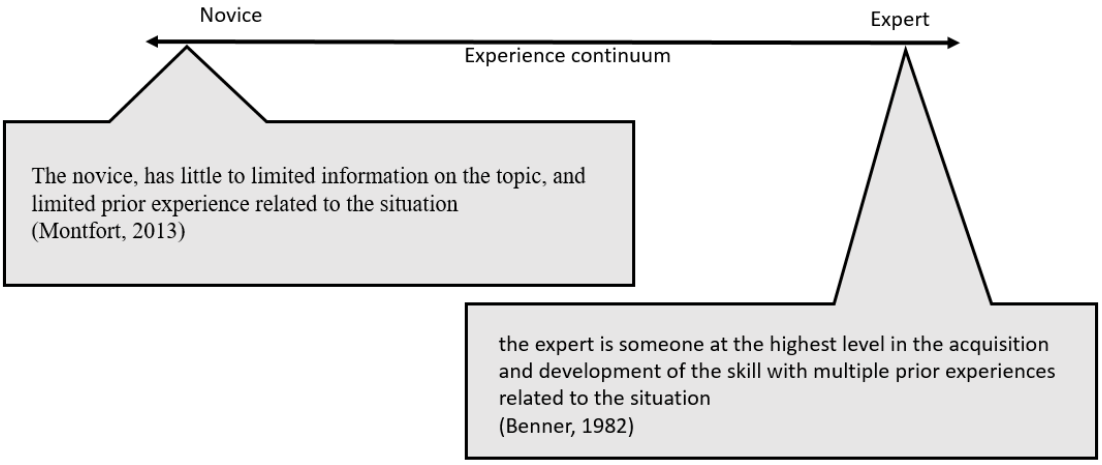
Similar approaches to automated text analytics usually involve use of computer-based algorithms employing Natural Language Processing techniques to extract information from large text-based data. Natural Language Processing may be understood as the use of computer manipulation of natural languages (Bird, Klein, & Loper, 2009). In research studies using Natural Language Processing, words or phrases are first converted to a numeric format for easier processing by the machine. Herzog and Benoit (2013) explain that to analyze text as data, documents and features must first be converted into a structured, numerical format. This structured form is then used to extract trends and patterns based on the features of interest (such as word or phrase frequencies, and noun-verb use).

Extracting trends and patterns from text-based data may be useful to researchers working in -related contexts. Text analytics using automated systems developed through use of Natural Language Processing and machine learning based techniques have been used for a variety of purposes in educational contexts from profiling student interactions in threaded discussions using speech act classifiers (Ravi & Kim, 2007) to an automated system which can automatically generate multiple choice questions on English grammar and vocabulary based from online news articles Hoshino and Nakagawa (2005). More recent publications include classifying exam questions using Bloom's Taxonomy (Jayakodi, et al. 2016) and keyword extraction from educational video transcripts through

use of Natural Language Processing techniques (Shukla & Kakkar, 2016). Extending computational linguistics through use of Natural Language Processing for engineering classrooms, Variawa et al. (2013) described an automated method to develop course-specific vocabulary and described how their model successfully identified domain-specific terms on engineering exams. Natural Language Processing techniques have thus been used for analyzing and extracting information from textual datasets in Education (Bhaduri, in preparation).

2.2 Novice-led approach to qualitative analysis

A novice can be identified as a person who 'has no experience with the situations in which they are expected to perform tasks' (Benner, 1982). In terms of skill acquisition on the basis of proficiency, the novice is at the lowest level in the acquisition and development of a skill. The novice, thus has little to limited information on the topic, and limited prior experience related to the situation (Montfort, 2013). In contrast, the expert is someone at the highest level in the acquisition and development of the skill with multiple prior experiences related to the situation (Benner, 1982). These varying levels of experiences with a situation or a skill, may be thought of as a continuum. In Figure 1 below, I depict the experience continuum with focus on the two extremes of the spectrum: novice and expert.



Manuscript Three. Fig. 1 Showing the experience continuum with novice and experts.

For example, limited background and expertise in the research topic of engineering student career decisions, for the scope of this research, helped me identify myself as a novice researcher. Further, and related more directly to the study itself in terms of data collection and analysis, in this research, I identify as a novice because I did not contribute to data collection or instrument development. I was the only novice on the PEPS team, in comparison to my advisor, a member of the PEPS team, who has more than 10 years of experience analyzing data related to career choices processes and may be considered an expert in the field. I do however recognize that, even though I am a novice in terms of knowledge of the existing models in literature on career preparedness among undergraduate engineering students, I am myself an engineer and have had to make several career-related decisions myself

Extending this definition of novices and experts to conducting qualitative research, we can understand the rationale presented by Montfort, et al. (2013) who described novice-

led paired thematic analysis. For Montfort, et al. novices differ from experts in that they lack shared assumptions as a result of being sufficiently unfamiliar with the content area. For example, Montfort, et al. describe how novices, while analyzing interview transcripts, may have some difficulty in interpreting the interview questions themselves, since they are sufficiently unaware of the literature in the subject area. An expert working with the novice can help the novice delve deeper into the subject area through multiple sessions of dialogue on the dataset where the novice may ask questions related to the data. Novice-led paired thematic analysis as described by Montfort, et al. (2013) thus make use of novice's unique approach to a given dataset which may be able to account for expert blind spots, while also making full use of experts' modes of thinking about the content matter.

The process described by Montfort, et al (2013) highlights how the novice needs to spend a lot of time and effort to familiarize themselves with the dataset. In dealing with large qualitative datasets, novice researchers may end up spending inordinate amounts of time in just reading through the transcripts. In elaborating on the coding process for exploratory qualitative research, Tesch (1990) as cited in Creswell (2014) recommend reading all transcripts carefully, and then clustering similar topics across all participants. For large qualitative datasets, reading through and iteratively making meaning of all of the transcripts in relation to one another may also be time-consuming for novice researchers. Transcripts from one-on-one interviews lasting for forty-five minutes to an hour are large, with the possibility of a transcript of a single interview to be 10-15 pages long. In addition to time intensiveness of the process, novice researchers may not know

where to begin while clustering for topics. For such large qualitative datasets, such as those involving excerpts from interview transcripts, Natural Language Processing techniques can help the novice by providing them with a quick overview of the dataset. For the PEPS project, Natural Language Processing techniques helped me as a novice researcher to familiarize myself with the dataset through looking at word clusters, word frequencies, and results of exploratory topic modeling. Analyzing results of these exploratory Natural Language Processing techniques led to observations about the dataset that I then conveyed to the larger PEPS team, comprising experts in the content area of engineering student career choices. The experts were able to use this information to streamline their initial objectives as well as refine their future directions based on the recommendations of my study.

2.3 Using Natural Language Processing techniques in a novice-led exploratory approach to PEPS dataset

The PEPS project, as introduced in an earlier section, was designed as a longitudinal study focusing on a three-year period from students' junior years into their early career years after earning their bachelor's degree. Currently, retention of engineers has been found to be a challenge faced by engineering educators, with fewer than 10 percent of US college students pursuing engineering degrees (Chen, 2013), and among those graduating with a degree in engineering less than half choosing to pursue engineering as their career

(National Science Board, 2010). Insights from the PEPS study may thus be useful for supporting diverse and qualified engineers in the field.

A part of the PEPS study involved interviewing junior and senior engineering students to understand their career choices and gauge their level of career preparedness. As introduced earlier, I was a part of the team analyzing the qualitative dataset, and contributed as a novice researcher. In the following sections I elaborate on how Natural Language Processing techniques were specifically incorporated in my novice-led exploratory engagement with excerpts from the interview transcripts, through a detailed discussion of the methods used in the study and a description of my interactions with the experts on the team.

3. Methods

In this section I will present the various Natural Language Processing techniques that helped in the exploratory data analysis of excerpts from student interviews related to career preparedness. Before I begin the describing the techniques, I first will present an overview of the dataset itself, including a description of the sites from which the data was collected. The sections on Data Collection and Dataset Overview will help provide an idea of the volume of the dataset, and will further motivate use of Natural Language Processing techniques to interact with it.

3.1 Data Collection

Brunhaver, et al. (2017) elaborate how the parent PEPS study takes a national perspective to better understand the issue of engineering students' career decision-making and preparedness. In order to answer the particular research questions driving my study which was part of the larger PEPS study, I used excerpts from interviews conducted with junior and senior engineering students at six partner schools as part of the PEPS study.

3.1.1 Site

The six sites selected for PEPS (ECOM, ERES, MPRI, MPUB, WPRI, WPUB) represent three geographic regions across the country including Eastern, Midwest and Western. The characteristics of each of these six sites is described in Table 1 below.

Manuscript Three. Table 1 Characteristics of six sites part of the PEPS study

Site	School	Control	UG Focus
ERES	Eastern Residential	Public	Arts and Sciences, Professions
ECOM	Eastern Commuter	Public	Arts and Sciences, Professions
MPRI	Middle Private	Private	Engineering
MPUB	Middle Public	Public	Professions plus arts & sciences
WPRI	Western Private	Private	Arts & Sciences
WPUB	Western Public	Public	Engineering

The sites described in the Table 1 are diverse in terms of the control (i.e., Public or Private), the focus of their undergraduate degree programs (i.e., Arts and Sciences, or

Engineering), geographic location in the United States (i.e., eastern, western, or in middle USA), as well as diversity of the student bodies (e.g., predominantly white males at MPRI versus higher enrolment of students identifying as Hispanic at WPUB). Brunhaver, et al. (2017) describe how these sites were purposively sampled for their 'geographic, institutional, and student body diversity', to facilitate examination of multiple factors related to both personal and context that may affect engineering students' career choices.

3.1.2 Population.

A total of 61 individuals were interviewed, across the six sites identified above, using a semi-structured interview protocol as part of the larger study to understand professional engineering pathways. Out of the 61 individuals interviewed, 38 were male and 23 were females, 22 were juniors while 39 were seniors in engineering programs. Thirty-nine self-identified as white/Caucasian, nine as Asian, and the remainder as a combination of underrepresented in engineering race/ ethnicities. Out of those interviewed, 57 were Mechanical Engineers while four were Chemical Engineers. All of the participants who were interviewed consented to the research through multi-institution approved IRB protocols.

3.2 Dataset Overview

Data analysis was conducted on a set of excerpts from interview data collected using a semi-structured interview protocol. The interviews averaged 43 minutes and focused on students' knowledge, skills and abilities related to getting their first position after

graduation (includes graduate school). Similar but separate protocols were developed for juniors and seniors anticipating that they would be at different points in securing first jobs after graduation. After initial questions seeking consent, and to establish ease of conversation through ice-breakers, the students were asked to provide examples about the knowledge, skills and abilities they thought they had or lacked regarding both on-the-job and job acquisition, and reflect on the extent of their career preparedness. Both protocols were divided into three main parts: (1) Career choice options and confidence (e.g., Have you thought about your plans for after you graduate? What types of jobs are you planning on and what factors are influencing your job choices?), (2) Getting the job (e.g., What knowledge, skills and abilities do you believe to be most important in obtaining your first job; how do these knowledge, skills and abilities interrelate?), and (3) Reflection (e.g., What is your expectancy of success of landing that first job you want? How important to you is it to get that job?).

3.2.1 Manual Extraction of Excerpts and Coding

The interviews were transcribed and uploaded onto Dedoose software for analysis. An expert, familiar with literature on engineering students' career preparedness then randomly selected about 50% of the data, i.e., 34 transcripts. For these 34 transcripts, the researcher assigned codes to segments of the interviews which they believed related to the student describing their career preparedness. The expert researcher used Expectancy Value Theory (Wigfield & Eccles, 2000) to describe participant preparedness as 'the belief in themselves'. As an initial start to the data analysis, the expert researcher then

isolated all excerpts, which were tagged as related to preparedness, and then assigned levels for this preparedness as High, Neutral, or Low. The expert explained these levels as High for when a participant describes having a job offer, Neutral as indicating that the person has taken steps to indicate preparedness but does not explicitly claim preparedness, and Low where there was a lack of experience or action that would qualify the participant as prepared for the career.

3.2.2 Dataset Used for Exploratory Analysis

The excerpts from the interviews, and the corresponding code for high, neutral or low related to career preparedness as evident from the particular excerpts were then downloaded from Dedoose as an Excel spreadsheet and presented to me for my analysis. I conducted my exploratory analysis on 296 excerpts. Of these 296 excerpts, 94 excerpts belonged to female participants and 193 excerpts were from male participants, while 9 excerpts were from participants who did not wish to disclose their sex. 73 of the excerpts were from participants in their junior year of engineering at the time of the interview, while 223 excerpts were from participants in their senior year of engineering at the time of the interview. Based on the codes accompanying the excerpts, 203 excerpts were tagged as High for career preparedness. Out of these 203 High excerpts, 77% of the excerpts were from participants who were Seniors. Since these excerpts were from the interview transcripts of 34 participants, multiple excerpts could belong to single participants. However, I was not concerned with tracing the source participants for multiple responses, and conducted my analyses at the excerpts level, rather than the

participant level, thus treating each excerpt and its corresponding label for career preparedness as a distinct data-point.

3.3 Data Analysis

Over the course of interactions with the expert analysis team, I realized an iterative process of "analyze-feedback-analyze" would best help their exploratory analysis on student preparedness for their first position after graduation. My initial analysis evolved around their preparedness ratings (High, Low, and Neutral); however, my results helped verify their concerns that the ratings were not robust. Based on a collaborative discussion on my results, I next provided data on word usage frequencies, word clusters, and patterns with assumed key independent variables (participant sex and grade level). These data were valuable in aiding the PEPS expert coders with insight into a new coding scheme for assessing students' levels of preparedness.

Though the initial purpose of my analysis was to understand if rankings of high, neutral and low were meaningful with regard to the types of words, phrases and sentiments of participants. However, an automated classifier performed poorly on this classification task and was unable to classify with high accuracy the excerpts on the basis of the level of career preparedness originally assigned by the experts. As stated above, we realized the nuances of the rankings were not well understood. Therefore, we next decided to instead focus on a more exploratory analysis to first understand the dataset in terms of the words comprised and the breadth of topics these words may be addressing.

These exploratory analyses were carried out using Natural Language Processing techniques to analyze word frequencies, word clusters and model the words to find topics present in the dataset. Findings were iteratively shared with the research team to direct further manual coding approaches.

3.3.1 Data Pre-processing

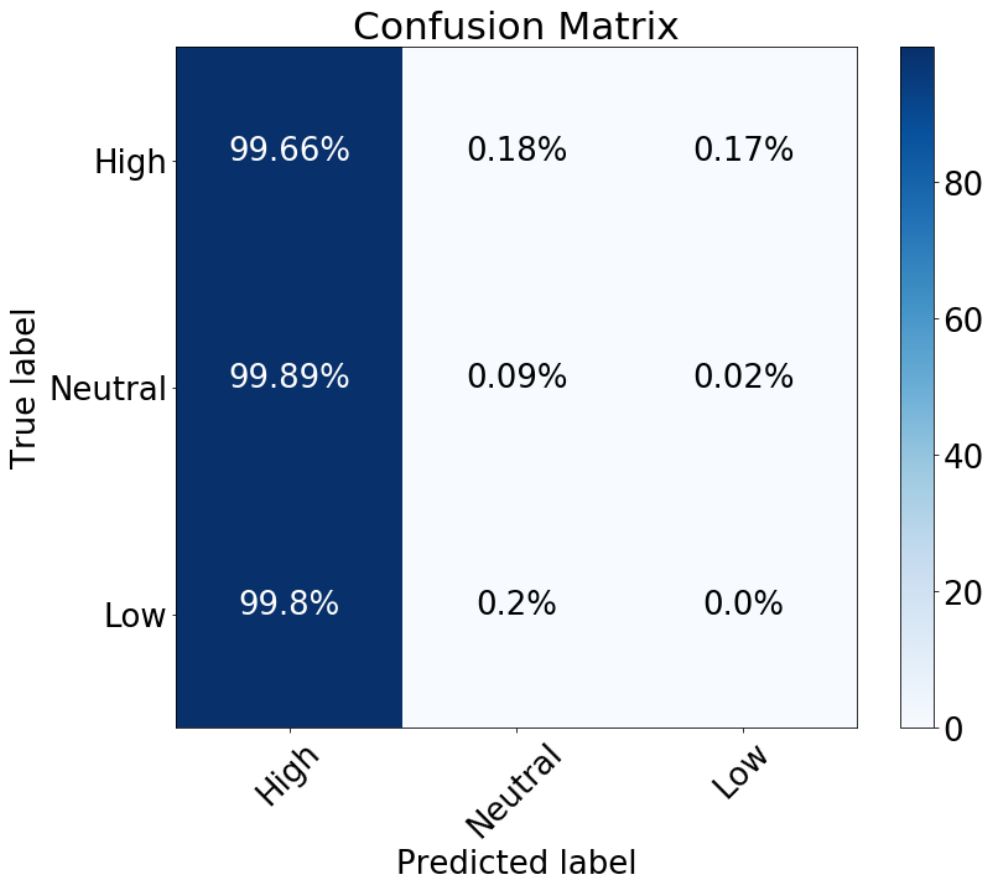
The pre-processing step conducted in this research was similar to the one in the work by Bhaduri (in preparation). Thus, I first removed all of the punctuation marks from the document and then converted all the words to lower case. I then stemmed the words to their root and removed English stop words.

3.3.2 Attempts at Automated Classification of Excerpts

The original direction of this research was to enable automated classification of the excerpts to understand differences among excerpts coded as High, Low or Neutral for career preparedness by an expert. The excerpts, following pre-processing, along with the corresponding expert assigned level indicating level of career preparedness were collected and then used for developing and testing an automated classifier. NLTK (Natural Language Toolkit) which is a suite of open source Python modules (citation) was used for the classifier development and testing.

The automated classifier used was a replication of the one used by Bhaduri (in preparation) to classify engineering student responses into High, Medium or Low based on metacognitive levels. I first partitioned the dataset into an 80-20 split for the purpose of classification. I then trained the classifier on 80% of the data, and tested on the

remaining 20%. While in the classification for the metacognition levels the automated classifier was able to classify with an accuracy of 66%, for this dataset the classifier failed at the classification task by incorrectly classifying most responses in the test set as High. As can be seen in the confusion matrix below (Figure 2), the predicted labels (across rows) seemed to predict most of the excerpts as High. Reading along the diagonal, we see that the High responses were correctly classified 99.66% of the times, while the Neutral and Low responses were correctly classified only 0.1% and 0% of the times. These percentages are based on average classifier performances for a Logistic Regression based classifier averaged over 1000 runs. For each run the system randomly selected training and testing samples based on the 80-20 stipulation. Thus, the machine was unable to distinguish the excerpts based on the levels for career preparedness assigned by the expert.



Manuscript Three. Fig. 2 Results of Classifier averaged over 1000 runs.

3.3.3 Exploratory Data Analysis

The results from the automated classification of excerpts led the research team to rethink the analysis process. As the novice researcher, I explained to the team how the classification scheme did not seem to work. This led to a change in the direction of my approach to analysis. I began looking more carefully at what constituted preparedness. To understand career preparedness from the dataset I first looked at word frequencies,

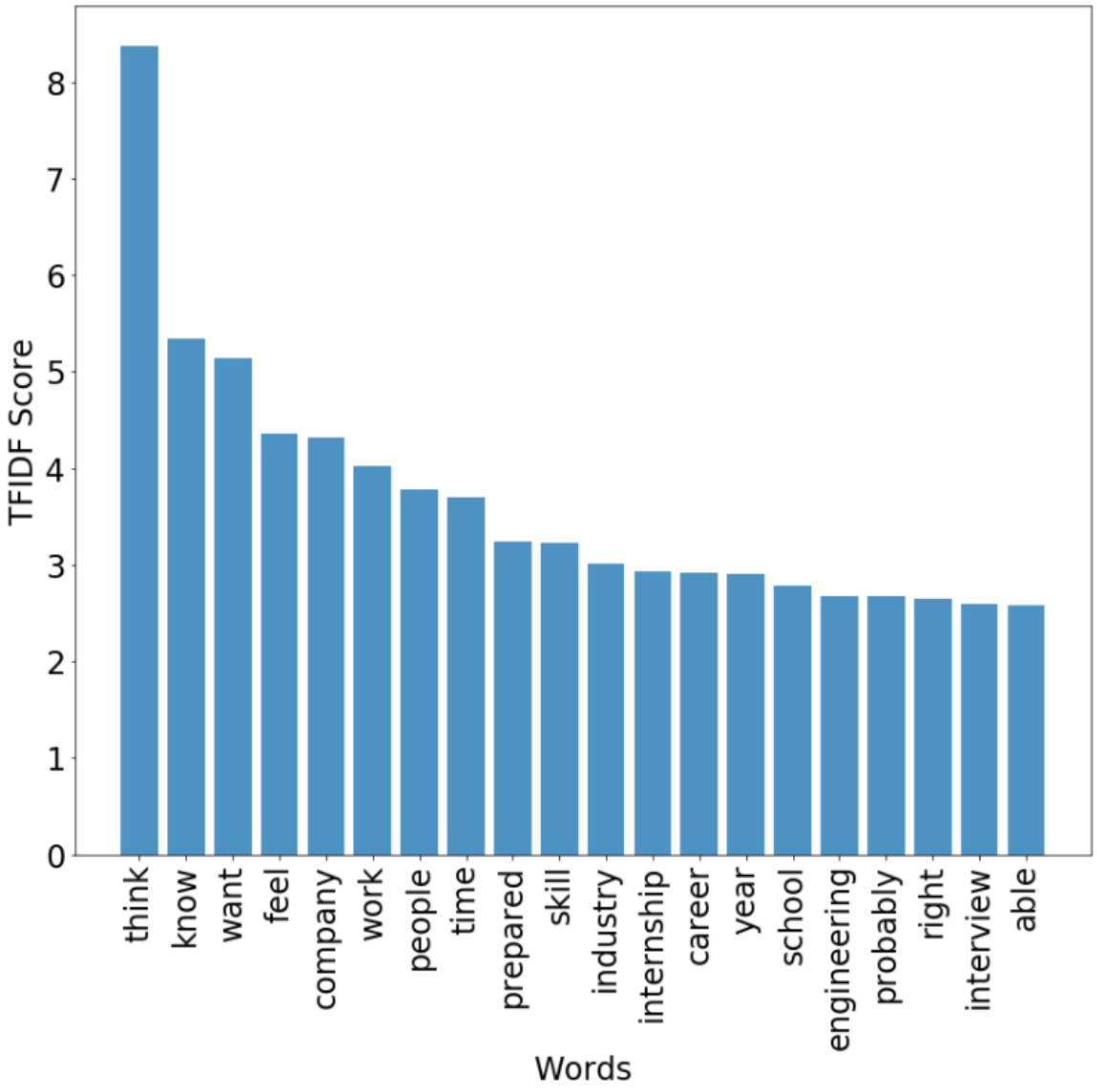
followed by an approach to automatically cluster these words based on semantic similarities, and finally tried to look for topics in the dataset

3.3.3.1 Word Frequencies

The first step in the exploratory analysis was to understand the words that were part of the dataset. An overview of the datasets was presented to the research team using common words that the participants used to describe their career preparedness. To isolate the frequently used words across the excerpts, I conducted a tf-idf based analysis. In analyzing large text corpora, tf-idf is a popular method which goes beyond the word frequency to not only compute frequency for the word in a particular document, but also multiplies this frequency by the inverse document frequency which leads to lower tf-idf scores for words such as articles (a, an, the) which occur in large instances in both single documents, and in the entire corpora (Shi, Xu, & Yang, 2009; Variawa et al., 2013). Variawa (2014) further described the computation of tf-idf scores:

> "The TF is a number determined by counting of occurrences of a particular word, and dividing that number by the total number of words in the target document: as such, it is a measure of frequency. The IDF is a measure of how important a particular term is within a set of documents, and is calculated by dividing the total number of documents by the number of documents in the set which contain that term, and then takes its logarithm. The TF-IDF formula multiplies these together and attaches the resulting score to each unique word in the target document." (pg. 204)

I then used the tf-idf scores to generate the top twenty words in this dataset The Figure 3 below shows the graph of the top twenty words in the dataset with the highest tf-idf scores.



Manuscript Three. Fig. 3 Top 20 words based on TF-IDF scores of all of the words in the dataset

As can be seen from the graph, the excerpts included words like think, company, know, feel, work, and people, among others. I presented my results from the analysis of word occurrences and scores from the tf-idf to the research team. Based on the results we agreed that the words were indicative of contributing factors to the career preparedness of the participants interviewed. This step also helped triangulate the manual coding process by the experts since the excerpts themselves were chosen based on them being indicative of students' career preparedness, which was also a finding of my analysis.

Knowledge of the high tf-idf words helped me guide a discussion with the research team on the fact that most participants repeat the use of certain words that have also been used by other participants. The research team saw value in the findings especially in observing words like company, work, people, interview and internship; as these could provide clues to the factors that contribute to preparedness for the participants as well as an awareness of people thinking about and knowing about these items. Based on our discussions, the expert researchers in the team then asked me to isolate nouns with the highest tf-idf scores, as well as investigate if there were differences in the words based on their tf-idf scores for excerpts from male versus female participants, and Junior versus Senior participants. A more thorough discussion of the findings are presented in the Findings section. Again, the purpose was not to declare findings as similarities and differences in the groups based on word frequencies alone but rather to provide a starting point for the team to engage in a richer analysis of the interviews.

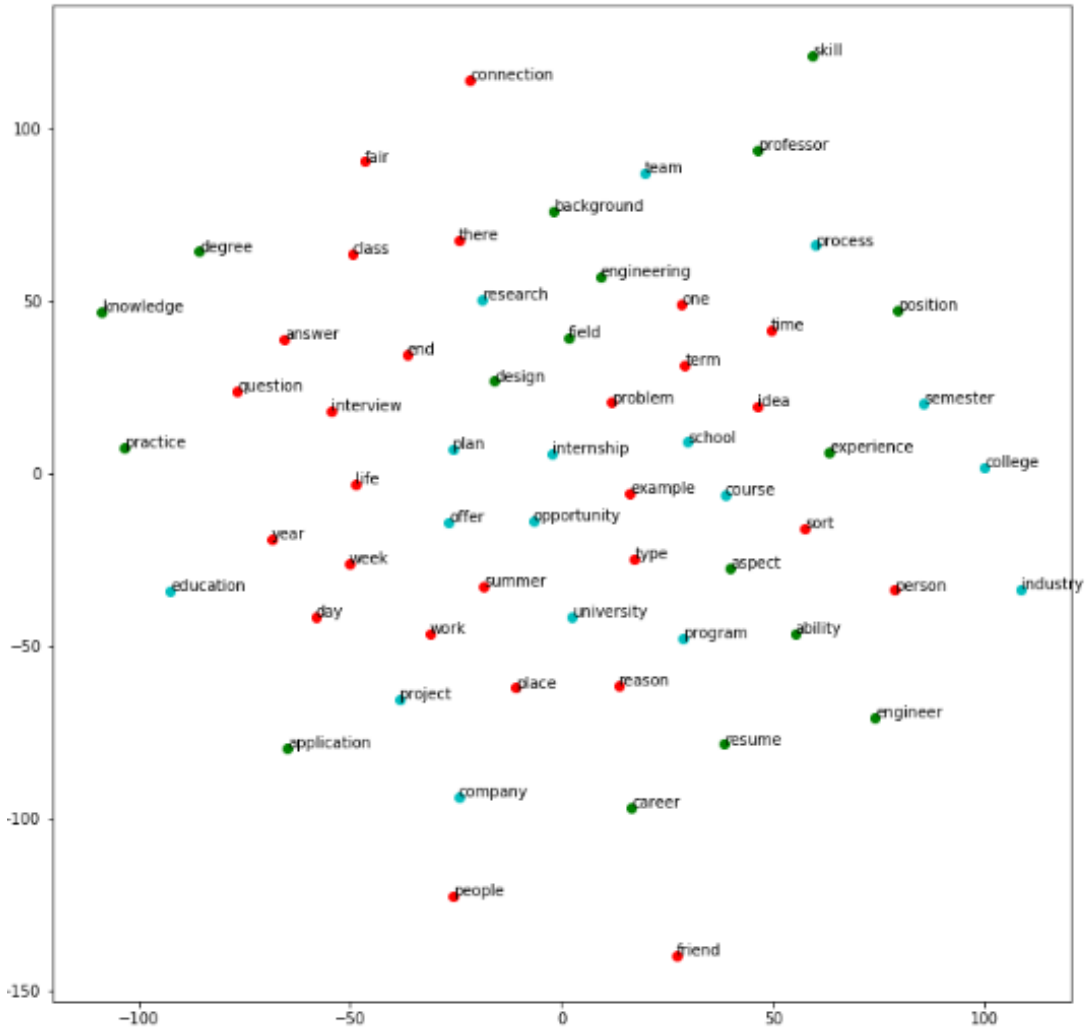
3.3.3.2 Clustering

Following the above step of understanding the distribution of words in the dataset on the basis of their tf-idf scores, the team was interested in exploring if the nouns that were part of the dataset would cluster together into meaningful categories. To conduct clustering, I used GloVe word embeddings, an unsupervised method to first obtain vector representations for all the nouns in the dataset. GloVe was developed by researchers at Stanford University (Pennington, et al. 2014). By this approach, words in a dataset are converted to numerical vectors which then help in the task of forming clusters. This approach of converting words to vectors has been used in other studies for Engineering Education data, e.g., Bhaduri and Roy (2017) conducted a mapping of words in mission statements of universities by first using a word2vec embedding (Mikolov, et. al. 2013) to convert the words in their data to vector forms. Word2vec, similar to GloVe is a word embedding which was developed by researchers at Google (Mikolov, et al. 2013).

Elaborating on how vector representations of words are useful for clustering, Mikolov, et al. (2013) describe how 'similarity of word representations go beyond simple syntactic regularities'. For example: the following vector arithmetic expression for individual words within the parenthesis "vector (" King") – vector (" Man") + vector (" Woman") results in a vector that is closest to the vector representation of the word "Queen". Once each noun was converted to a vector using the GloVE word embeddings, these were clustered using NLTK's in-built k-means clustering algorithm to form three clusters. These clusters were completed based on the cosine distances of the words, from one another in an n-dimensional space. Cosine measures calculate the similarity between two objects through the cosine angle between the feature vectors in n-dimensional space. I

used the in-built NLTK functions for calculating the cosine measures. The final step was to project the nouns onto a 2-D plot, presented in Figure 4, to have a visual mapping of the nouns part of the three clusters. In Figure 3, we can see how nouns from my excerpt dataset are clustered into three different sets based on the cosine distances between the word vectors in an n-dimensional space. It can be seen that nouns like “interview”, “question”, and “answer” clustered together (i.e., nouns depicted by the Red dots), while nouns like “internship”, “opportunity” and “offer” formed a different cluster (i.e., nouns depicted by the Cyan dots), and nouns like “background”, “engineering”, “field” and “design” formed the third cluster (i.e., nouns depicted by the Green dots). These three different clusters tabulated in Table 3 (pg.108) helped me think about the three broad ways in which the words of the dataset could be grouped.

Manuscript Three. Fig. 4 Words represented in 2D space as part of one of three clusters (Red, Cyan or Green)



Drawing analogies to coding techniques, this clustering technique can be thought of as similar to forming categories from existing fragmented open codes, such as the procedure followed in Grounded Theory approaches to analysis. However, unlike conventional coding techniques for qualitative datasets, these clustering techniques were fast to implement and thus time and resource efficient in providing an overview of the dataset Recall that the total data set contains 62 interviews with an average of 43 minutes

duration for a total of over 2500 minutes of data. Conventional coding techniques would likely take up weeks of time for analysis for initial coding, and weeks for clustering. In contrast, the clustering generated by the algorithm was obtained within a few seconds. I used the results of this cluster analysis to inform my understanding of the types of words that were present in this dataset Details on the interpretations of the individual clusters are presented in the Findings section.

3.3.3.2 Topic Modeling using LDA

The final step in the exploratory analysis process was that of topic modeling. I used a method called as Latent Dirichlet Allocation (LDA). LDA can be understood as a probabilistic model for a textual dataset The idea behind LDA based topic modeling is that a textual dataset can be represented as a random mixture of topics, where each topic is characterized by a distribution of words (Blei, Ng and Jordan, 2003). Maskeri, Sarkar and Heafield (2008) elaborate on how LDA can be used for a document to identify: (1) a set of topics in a document, (2) the words associated with the topics, and (3) the mixture of these topics for the document. LDA has been used for topic modeling for large qualitative datasets. Savoy (2015) explains the working of LDA by elaborating how given a corpus (i.e., dataset), and a number of topics, the LDA returns for each topic, a list of words which occur in that topic. For example, Newman, Chemudugunta, Smyth and Steyvers (2006) used LDA for topic modeling of a dataset comprising 330000 news articles in the New York Times and found 400 topics such as Harry Potter, Basketball, and Holidays. They were then able to then represent each news article in their dataset as a mixture of these topics. Similarly, consider the excerpt dataset of this study. I used LDA

to find two topics that could be used to model the dataset Table 2 presents the two topics and the top ten words which are part of the two topics that were generated.

Manuscript Three. Table 2 Words in the two topics for my dataset found using LDA

Topic#0	think know feel interview probably year resume pretty got school
Topic#1	company class work want experience project important working time different

4. Findings

As introduced in an earlier section, the specific research question driving this study is: What insights can be gained from using Natural Language Processing techniques for exploratory novice-led analysis to inform opportunities for future analysis for a team of researchers working on understanding engineering student choices related to career preparedness?

Some of my key findings were that: (1) the words with high tf-idf scores provided insights into factors that affected student career preparedness, (2) clustering the nouns into three categories led to observable differences in the types of influencers that students described and which were tagged as nouns, and (3) there were no differences across groups for males versus females or Juniors versus Seniors in terms of the composition of the excerpts, however, some differences in word choices were observed across the two sets of groups. Consistent with the research question, these results are not intended to be

reported as findings by themselves but rather they informed further analysis of the data (as described in this section) to understand the patterns identified.

4.1 Insights from Overall Data

As part of the larger PEPS project, the expert team intended to analyze student's levels of preparedness for the process of career discovery in job acquisition. This required analyzing the transcripts for levels of preparedness related to learning about career options (career discovery phase), seeking a job (including learning about potential jobs, applying, interviewing, receiving offers), and accepting a job offer. The expert team's intended method was to code individual segments related to preparedness throughout each interview using a constant-comparative method (Fram, 2013; Charmaz, 2014) for individual codes. The use of a constant-comparative method is thorough as it requires the results of each coded transcript to be compared back to previous transcripts to ensure consistency of codes and coding. The disadvantage of this method is it is time consuming when used with a large dataset In addition to coding the excerpts as preparedness related, the expert team also assigned the level of preparedness. Due to the paucity of preparedness data related to undergraduate students obtaining their first position after graduation, the criteria to discern different levels of preparedness needed to emerge.

Because the original analysis of preparedness resulted in a majority of excerpts being coded as High, additional information on what, if any, differences could be noted to help distinguish High, Low, or Neutral classifications was sought by the research team.

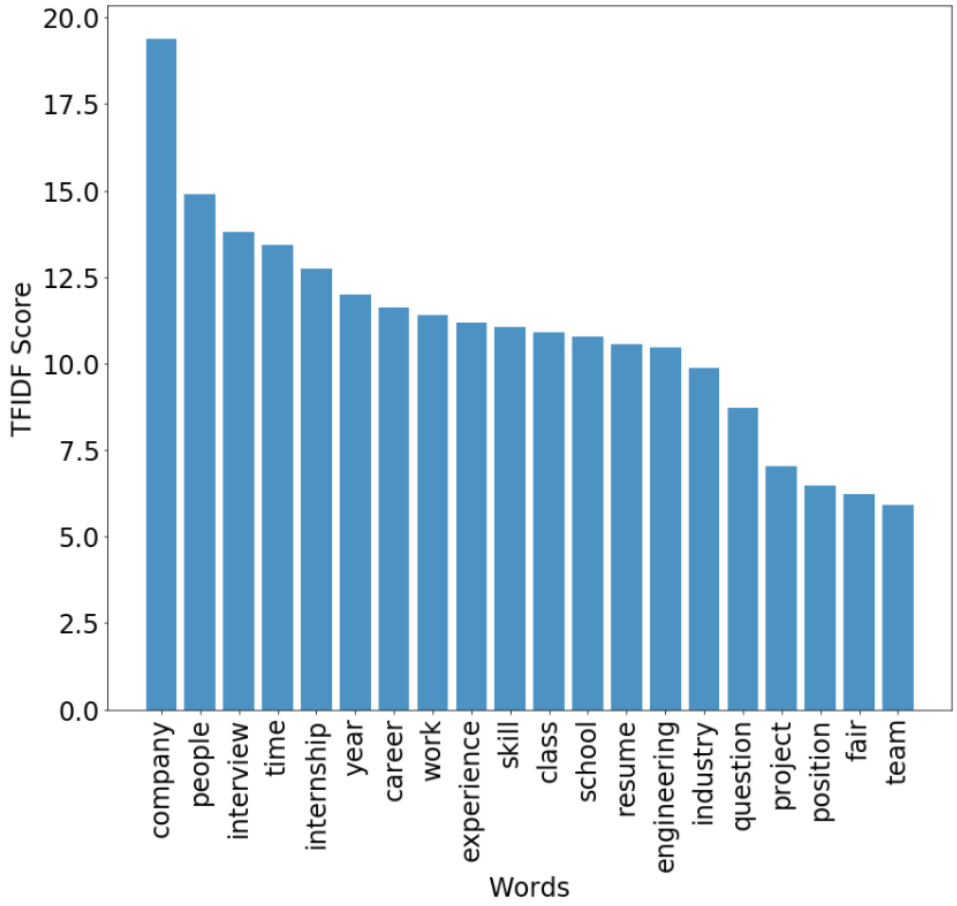
Information provided by the results of classification accuracy of the Logistic Regression based automated classifier revealed that the initial discriminators used to classify High, Low, or Neutral were unlikely to provide clear distinctions. Fortunately, the most commonly used words provided insight on a way for the experts to direct their coding by considering what the words "think" and "know" may be portraying. As a result, the team asked for additional data analysis on the use of top nouns and word patterns to see if they too could help in understanding the most common themes in the data.

A Logistic Regression based automated classifier was unable to distinguish the excerpts based on the levels for career preparedness assigned by the expert. Thus, the expert team redirected their efforts from analyzing a combination of quantity and variety of preparedness "items" (i.e., the Highs, Lows, Neutrals) to analyzing the excerpts based on Peterson, Sampson et al (1991) pyramid of information processing domains. Within this framework, knowing about self and knowing about options are base items, when combined they help to form decision-making skills. Finally, metacognition is at the top of the pyramid. Likewise, my Natural Language Processing based cluster analysis enabled words to be grouped into the three clusters providing insight into which phases of career decision making appeared to be most salient to the participants. For example, the cluster comprising words such as "interview", "question", and "answer" could be offering insights into words used by participants in describing actions during the job acquisition phase; similarly, for possibly the phases of career discovery or learning about jobs, words such as "internship", "opportunity" and "offer" may be meaningful; and

finally, the cluster of words such as “background”, “engineering”, “field” and “design” may be insightful regarding career skills.

4.1.1 Potential factors contributing to career preparedness identified

As described in the Methods section, I generated tf-idf scores for each word in the dataset to identify potential factors contributing to career preparedness. I used the in-built NLTK Part of Speech tagger to tag all the nouns in the dataset and then generated the graph in Figure 4 based on the tf-idf scores of the nouns in the dataset.



Manuscript Three. Fig. 5 Graphing top 20 nouns based on TF-IDF scores for all nouns in the excerpts

The nouns in the dataset revealed some important stakeholders that the participants talked about while elaborating upon their career preparedness. For example, the noun with the highest tf-idf score was “company”, followed by “people”, “interview”, “time” and “internship” in descending order of the magnitudes of their tf-idf scores. For example, in the excerpt from a Male Junior, the participant reflects:

> “I think the most meaningful part of the job search world is definitely **interviews**. I think once you get your foot in the door, have an **interview** with a recruiter or a manager or whatever, I think that's where you really can understand what a **company** is like and you get to pick the brains of an insider, which you can't normally do.”

From this excerpt, we find how the word company is important to the participant as he describes the job search process. From this word, thus, it seems evident that multiple participants described their job preparedness in relation to the companies to which they were applying. Words such as “work”, “experience”, “skill” and “class” had similar tf-idf scores. Consider the word “skill”, a female senior participant elaborated:

> “One of the things I'm really interested in is biomechanics and surgical robotics. There I think my electronics and robotics skills are really important. Probably not as important are knowledge of the human body and anatomy. Usually they have some sort of medical director on the team or someone like that who would take care of those. Another would be mechanical design and mechanical systems design.”

Here, the participant elaborated on the skills that she deems necessary in relation to her career preparedness. Thus, a feasible scope for future research could be in terms of an

analysis of the skillsets that the participants elaborate upon as necessary for their career preparedness. Thus, these words and their corresponding tf-idf scores provided an overview of the plausible factors contributing to the career preparedness described by the participants.

4.1.2. Insights from Clustering

The clustering process described in the Methods section was conducted using k-means clustering, to cluster the words in the corpus into three main clusters based on the cosine distances of the word vectors from each other. I have described previously how these word vectors were formed using GloVE word embeddings (Pennington, Socher and Manning, 2014). In Table 3 I isolated the words from each cluster (shown in Figure 3 with cyan, green and red colors) and present in a tabulated form.

Manuscript Three. Table 3 Words part of the three separate clusters from the excerpt dataset

Cluster 1: Related to Personal Experiences (Red)	Cluster 2: Related to Knowledge/Skills (Green)	Cluster 3: Related to Opportunities (Cyan)
people	engineer	school
friend	career	internship
connection	experience	offer
fair	skill	company
interview	ability	industry
question	engineering	team
year	field	research
time	aspect	project
sort	knowledge	course
life	practice	plan
idea	degree	education
class	resume	college
there	application	opportunity
problem	professor	process
end	position	university
type	background	program
one	design	semester
place		
answer		
person		
summer		
term		

work
example
week
reason
day

In analyzing the words in the three clusters, I found that words in each category helped me think about career preparedness in terms of the participants' personal experiences, knowledge/skills, and opportunities/resources available. The first cluster seemed to comprise words related to personal experiences of the participants. Thus, there were words such as "friend", and "life". For example, a Male Senior participant reflected:

> "A lot of my friends, they had higher GPAs than me, so they scored jobs in October, which was good for them, but for me, it kind of hurt my confidence because I didn't have a job yet I think after that first career fair when I didn't have a job after October, it changed. I put in a lot more effort because I knew I'd have to put in more effort than my peers to get a job."

From this excerpt, it seems that the participant is elaborating on personal experiences related to career preparedness, explaining how the performance of his peers pushed him to work harder to secure a job. Within the career preparedness taxonomy, this relates to knowledge of self. Similarly, focusing on the word "life", I found a Male Senior participant reflect:

> "They just wanted me to go to college and get a degree in something that I wanted to do. They said that's the most important thing in life is to enjoy what you do for work, because you have to do it a lot."

This excerpt reflects another instance of a personal experience that shaped the participants' career preparedness. Thus, the words in the first cluster, for me seemed to indicate words that participants used to elaborate on personal experiences that influenced

them to think about what they wanted to do and what they enjoyed doing in life, which may have had an effect on their career preparedness and choices.

The words in the second cluster, seemed to indicate Knowledge/Skills related to career preparedness used by the participants. Words in this cluster apart from “skill” and “ability” included “engineering”, “career”, and “professor” among others. In going back to the excerpts for the individual words, I found that in using these words the participants usually described a skill or knowledge that they believed was important to acquiring a job, as well as resources for these skills, knowledge of career acquisition options within the career preparedness taxonomy.. For example, a Female Junior participant used the word “engineering” as she described her skillset in terms of career preparedness as”

> “I worry that I wouldn't be very good at if they ask technical engineering questions from a thermal class I took two years ago. I worry I wouldn't be able to remember that.”

This excerpt highlights how the participant believed that her abilities to answer engineering related questions were lacking, thus affecting her perception of her career preparedness. Similarly, another Female participant who was a Senior, elaborated on how she might prepare for a technical interview by consulting with a professor, the word “professor” being a part of the second cluster. Finally, the word “career” shows up in this cluster, and has been used in multiple ways in the excerpts. Students have spoken about the career centers in their universities who helped them build and critique resumes, or offered job postings, career has also been used in phrases such as “career plans post-graduation”. Thus, it can be seen how many words in the second cluster are words

indicating participants talking about concepts related to skills and knowledge in relation to their career preparedness.

The words in the final cluster seemed to relate to opportunities that the students elaborated upon in talking about their career preparedness. Words in this cluster included "opportunities", "internships", "offer" and "project" among others. For example, a Male Junior participant recounts:

> "My university offers quite a few help sessions and what not for resumes, or just career stuff in general. My internship company also did as well. They had a couple workshops that I went to over the summer, so I feel like I'm doing all right"

From this excerpt it seems like the participant is elaborating on opportunities that were available/offered to him. Similarly, multiple participants also used the word "offer" to describe their existing job offers, or internship offers. Thus, words in the third cluster seemed to indicate relation to broader opportunities available to the students which impacted their career preparedness.

4.2. Comparing Across Excerpts

Moving away from a general overview of the dataset, the second part of my exploratory analysis also looked at exploratory comparisons across excerpts on the basis of the demographic characteristics of the participants. Primarily, I compared excerpts from male participants to the ones from female participants. Similarly, I also compared excerpts from Junior participants to Senior participants. For the two sets of comparisons, I present my analysis for the similarities and differences in the word distributions based

on tf-idf scores, and the topic models that emerged. Again, the purpose was not to reveal statisitically significant differences between groups but rather to inform further analysis.

4.2.1 Comparing Males' Versus Females' Excerpts

Comparing the excerpts from transcripts of Male participants to those of Female participants I found both similarities and differences. The similarities in terms of the compositions of the excerpts from a Part of Speech proportion point of view, and a topic modeling perspective are highlighted in the Table 4 below.

Manuscript Three. Table 4 Similarities in compositions for Males and Females

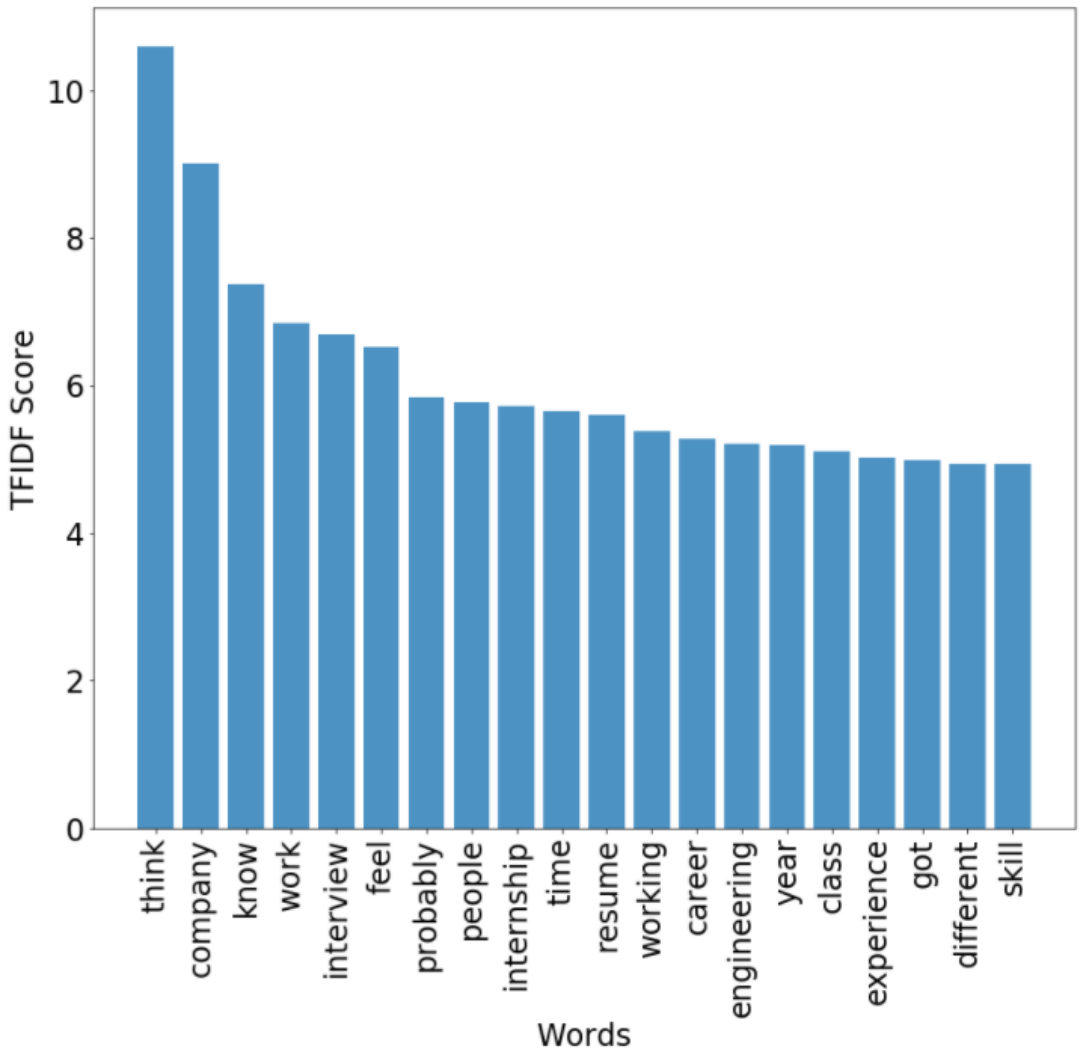
	Males	**Females**
Noun: Verb Ratio (indicating the average proportion of nouns and verbs for responses from each group, i.e., Male or Female)	17.31% **Noun** 20.05% **Verbs**	17.36% **Noun** 20.15% **Verbs**
Topic Modeling (indicating the topic composition for the two topics that were found using LDA)	**Topic #0:** 41.20% **Topic #1:** 58.79%	**Topic #0:** 40.19% **Topic #1**: 59.80%

Note: Here, Topic#0 comprised words: think, know, feel, interview, probably, year, resume, pretty got, school, and Topic #1 comprised words: company, class, work, want, experience, project, important, working, time, different.

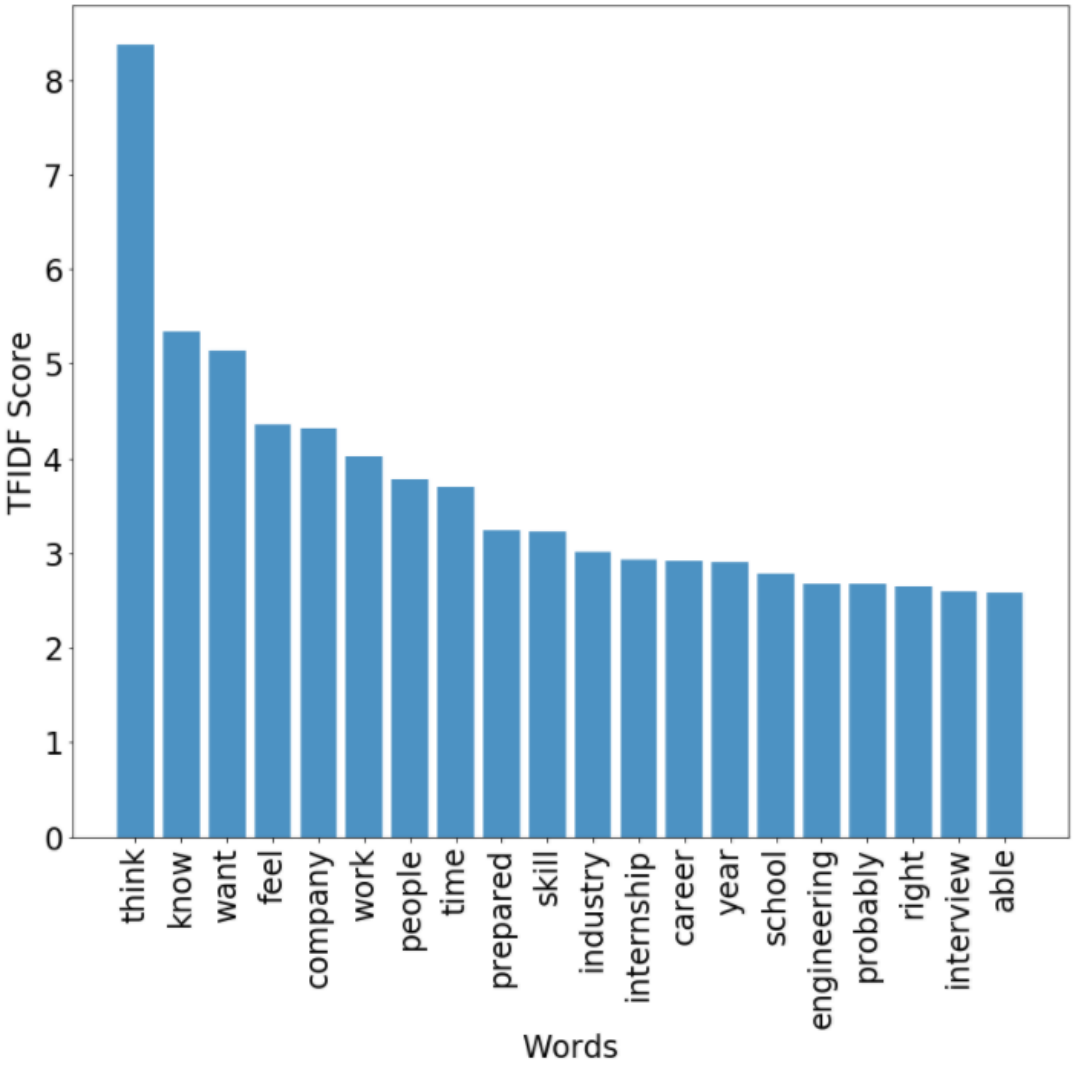
The similarities existed in terms of the compositions of the excerpts from both the two groups. I found that for the responses, the proportion of usage of parts of speeches were similar for both the groups. I also found that the topic modeling composition remained similar, showing a close to equal proportion of both the topics for both male and female datasets. The differences lay in terms of the specific words used, and the corresponding high tf-idf score words in both the groups. Although there were overlaps in the list of the top twenty high tf-idf words from both the groups, there was some difference in the order. I have presented the bar graphs comparing the tf-idfs in the Table 5 below. It can be seen that while the top word for both the groups was “think”, the next set of words for females were “know”, “want”, “feel”, “company”; while for males it was “company”, “know”, “work”, “interview”. The order of these words matter since it shows that the excerpts from the female transcripts had words such as wanting and feeling, were higher than company, while for males, company, work and interview made it to the top of the list.

Manuscript Three. Table 5 Comparing words with the highest TF-IDF scores for males versus females

TF-IDF for MALES



TF-IDF for FEMALES



4.2.2 Comparing Juniors' Versus Seniors' Excerpts

Comparing the excerpts from transcripts from Senior participants to those of Junior participants I found both similarities and differences. Once again, the similarities were in terms of the compositions of the excerpts from a Part of Speech proportion point of view, and a topic modeling perspective are highlighted in the Table 6 below.

Manuscript Three. Table 6 Similarities in compositions for Juniors and Seniors

	Juniors	**Seniors**
Noun: Verb Ratio (indicating the average proportion of nouns and verbs for responses from each group, i.e., Male or Female)	17.76% **Noun** 20.82% **Verbs**	17.17% **Noun** 19.86% **Verbs**
Topic Modeling (indicating the topic composition for the two topics that were found using LDA)	**Topic #0:** 38.22% **Topic #1:** 61.77%	**Topic #0:** 41.74% **Topic #1**: 58.25%

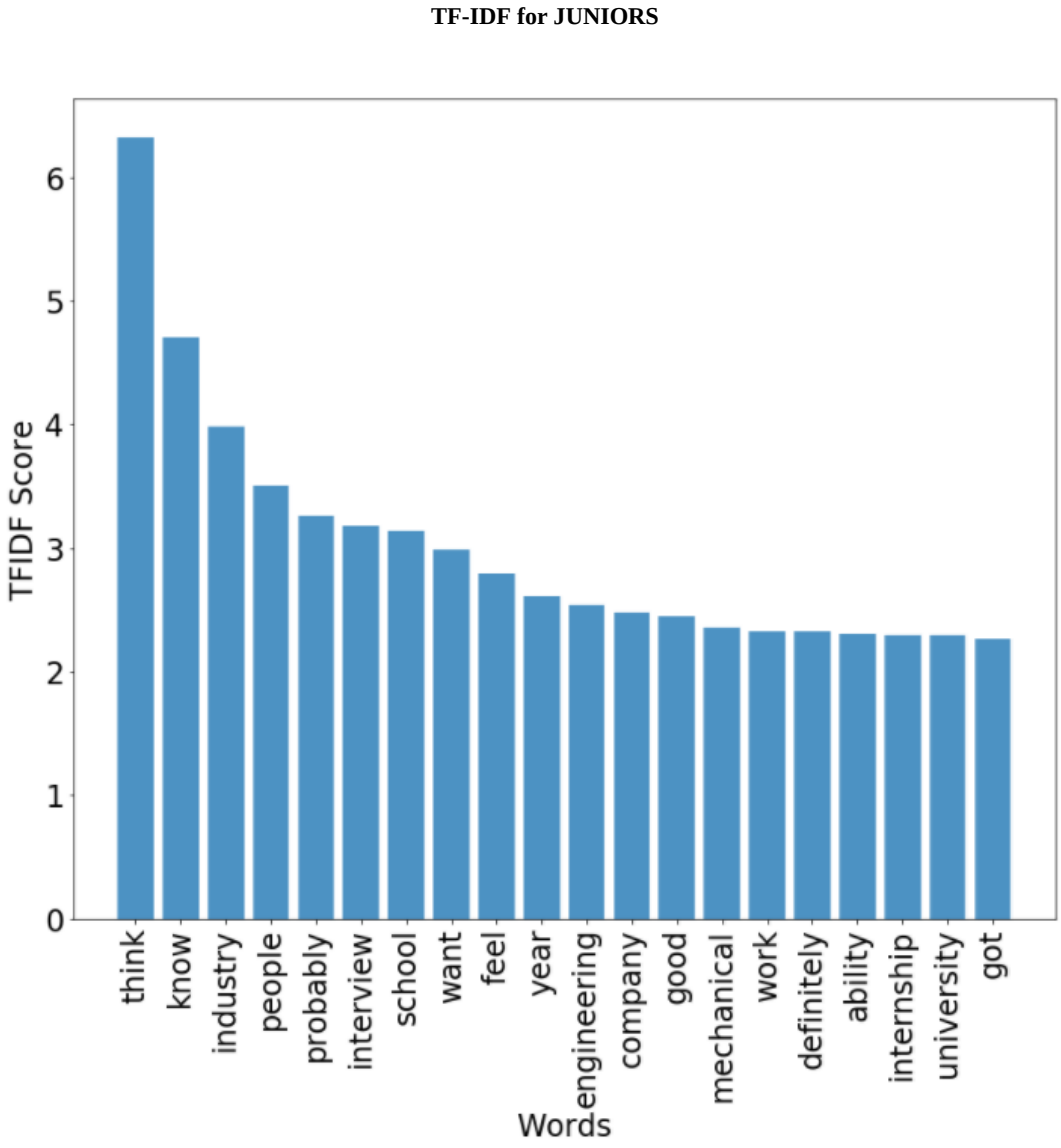
Note: Here, Topic#0 comprised words: think, know, feel, interview, probably, year, resume, pretty got, school, and Topic #1 comprised words: company, class, work, want, experience, project, important, working, time, different.

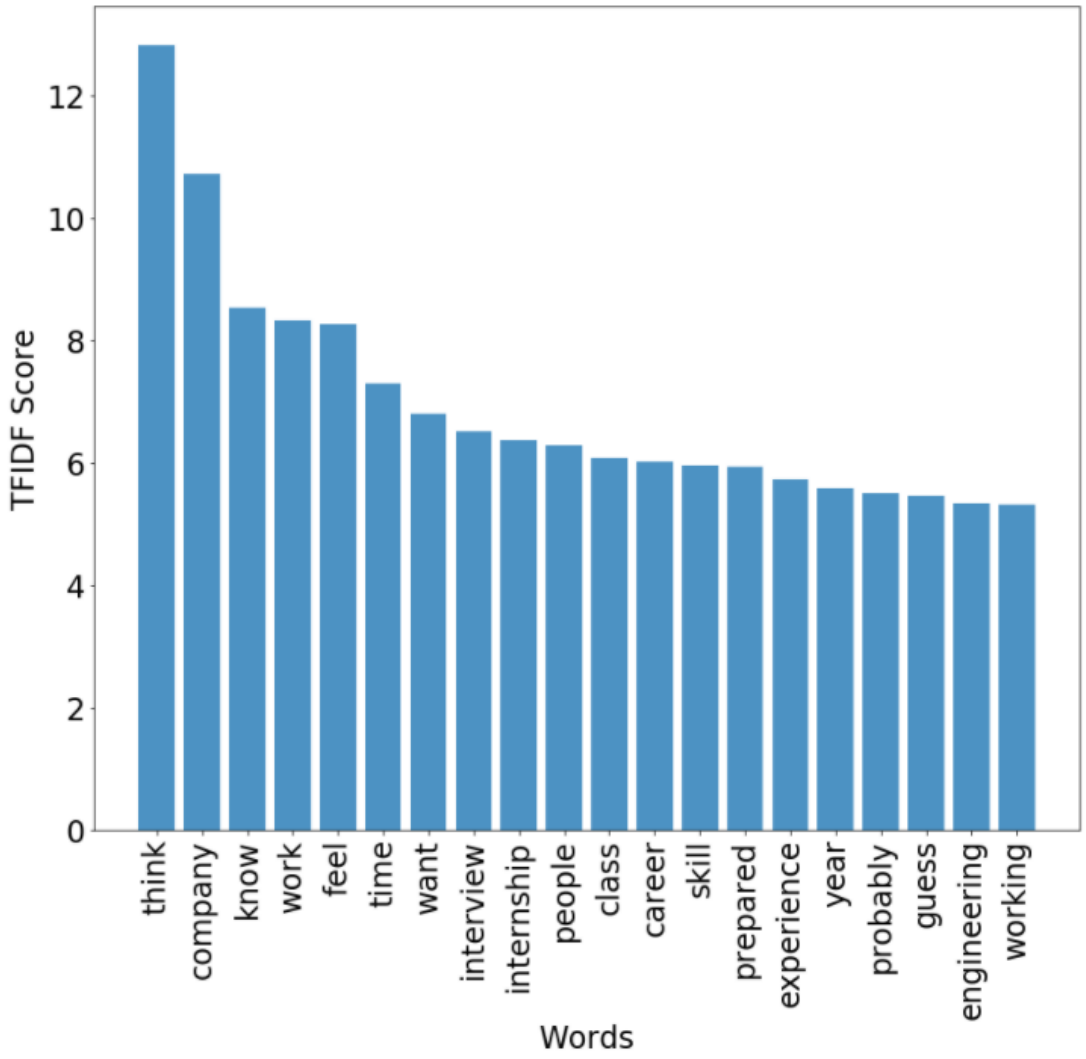
The similarities existed in terms of the compositions of the excerpts from both the two groups. I found that for the responses, the proportion of usage of parts of speeches were similar for both the groups. I also found that the topic modeling composition remained similar, showing a close to equal proportion of both the topics for both male and female datasets.

Once again, the key differences between the two groups was in terms of the order of the words based on their tf-idf scores (Table 7). Once again, “think” showed up in both groups with the highest tf-idf score. For Juniors, the next set of words were “know”, “industry”, “people”, “probably”; while for seniors it was “company”, “know”, “work”, “feel”. A note of caution in interpreting these results is that since a large proportion of the

excerpts from Seniors who were males, there may be an overlap in the tf-idf bar graphs from the previous sections.

Manuscript Three. Table 7 Comparing words with the highest TF-IDF scores for Juniors versus Seniors





5. Discussion

This study used a unique way to analyze interview transcripts, through helping a novice researcher make meaning of a large qualitative dataset with 292 excerpts from

student interviews, by using Natural Language Processing techniques. The results of this novice led analysis then informed my recommendations for future research directions to the larger research team comprising experts in engineering student career decisions. For example, I explained to the team how their initial classification scheme to classify responses as High, Low or Neutral in terms of career preparedness may not be working since results of a classification system was unable to accurately distinguish between the three labels. This led to a change in the direction of the team's approach to analysis of the dataset.

For my research, I next began looking more carefully at what constituted preparedness. Specifically, I looked at word frequencies, used part of speech tags to identify and then k-means to cluster the nouns into three distinct clusters, and finally used Latent Dirichlet Allocation (LDA) for topic modeling. These exploratory techniques were conducted by a novice researcher who was not aware of the constructs identified in literature related to career preparedness. Results of my study thus provided the team of qualitative researchers with information to streamline their initial objective of relating different levels of preparedness and with recommendations for future studies. Similar approaches to novice-led analysis can be used by other researchers in the field. Additionally, there are several possibilities for future work based on the results of my research which are elaborated upon in the following subsections.

5.1 Novice-led Analysis using Natural Language Processing Techniques

My research was influenced by the method of novice-led thematic analysis introduced by Montfort (2013) who has cited and how it has been used. A review of literature reveals another work by Scheponik, et. al. (2016) in which the authors describe the use of a novice led thematic analysis to address the issue of expert blind spot in analyzing cybersecurity courses related student reasoning. However, my study is unique in that it expands upon the novice led thematic analysis approach by facilitating analysis through use of Natural Language Processing by the novice researcher. Thus, I used Natural Language Processing techniques to familiarize myself, the novice researcher, with the dataset in a time and resource effective way. On the recommendation of the research team, I first worked on trying to automate classification of the dataset by mimicking the expert assigned labels of High, Low, and Neutral. A logistic regression based classifier failed to accurately classify the dataset. This led me to explain to the research team that their initial direction of classifying excerpts may not be working. As a research team we then decided to change the direction of the analysis on what career preparedness constituted, based on the given dataset Using Natural Language Processing tools, I next looked at word frequencies and tf-idf scores for words to understand word choices across the excerpts from the participants. I then used a k-means clustering technique to cluster the nouns in the dataset into three clusters based on their semantic similarities. Finally, using LDA I was able to model the dataset into two topics.

Thus, I found that using a Natural Language Processing based novice-led approach to exploratory analysis was particularly helpful for the specific context of this research study for two reasons: (1) since most other members of the research team were familiar

with the career preparedness literature, which made it helpful to have a novice conduct exploratory analysis, and provide unique insights about the dataset that may have escaped due to expert blindspots, and (2) given the multiple interests of the team, and the large size of the dataset, use of automated text analytical methods for the exploratory research proved insightful overview of the data, but at the same time was not resource or time intensive.

For example, based on the results of the Natural Language Processing techniques, the list of the most commonly used words may have provided insights to the experts to direct their coding by considering what the words “think” and “know” may be portraying. The three clusters of nouns that emerged from the dataset through the use of k-means clustering may also provide insights for possible directions for coding the dataset in the future. Thus, Natural Language Processing techniques were successfully able to provide the research team with an overview of the dataset without having to spend a lot of time and effort through constant comparative methods to understand patterns across the excerpts.

A similar approach can thus be adopted by researchers working in teams to analyze qualitative Engineering Education datasets, in similar contexts where there is a possibility that expert blind spots may necessitate bringing in a novice, and also large dataset size may require innovative approaches to minimize time spent on generating overviews from the dataset. An important lesson learned through this research exercise which may be relevant to other researchers embarking on studies using similar methods, is the iterative

and collaborative nature of the exploratory research. In my research, I consulted regularly with the larger research team to make meaning out of the information that the automated analysis revealed. Thus, through conducting this research, I found that the novice-led thematic analysis facilitated by Natural Language Processing is most effective when the speedy and resource-friendly analysis using algorithms is directed, interpreted, and understood through disciplinary expertise provided by the experts on the team.

5.2 Future Work

The results of my exploratory analysis can lead to several research directions for studies in Engineering Education. In the following sub-sections, I describe two of the more prominent set of directions for future work.

5.2.1 Factors contributing to career preparedness

One of the potential directions for future work may be to seek a deeper understanding of what students believed were important factors contributing to their career preparedness. From our exploratory analysis we found that the nouns seemed to cluster into three distinct categories, using the k-means clustering technique. I found that based on the similarity of the nouns, as per the embedding used the three meaningful clusters that emerged were: (1) related to personal experiences, (2) related to knowledge and skills, and (3) related to opportunities. In prior work (Carrico, Harris, Matusovich and Brunhaver, 2016), the research team had established the Knowledge, Skills and Abilities

that career services professionals believed to be critical for students to develop. One of the potential topics for future work could thus be to use qualitative analysis to find the knowledge and skills that students believed to be important to career preparedness and compare these to the ones that career services professionals deemed important. Another avenue for future qualitative research could be to understand the opportunities and resources available to the students. In this regard, the researchers could specifically look at influencers from the university (such as professors, career professionals) or personal (such as parents, friends, peers) who may be impacting the career preparedness of the students.

5.2.2 Differences due to demographic characteristics of engineering students

In my exploratory study, I did not find any significant differences in terms of part of speech distributions, or topic compositions among males versus females, or Juniors versus Seniors. I explain in the limitations section below how this lack of differences is possibly a result of the way in which data was collected, rather than indicative of absence of differences. I did however, find differences in the word choices, as indicated by the order of the words based on their tf-idf scores for each group. A number of factors could be contributing to these differences. Focusing on gender differences, we know from prior work that women are still largely under-represented in the undergraduate engineering classrooms (Bordogna, Fromm, & Ernst, 1995; Jamieson & Lohmann, 2009; Lichtenstein, Chen, Smith, & Maldonado, 2014; National Academy of Engineering, 2005). Tonso (2006) raised the issue of a chilly climate for women in engineering

education classrooms, through her publication elaborating on engineering teams. Similarly, Lichtenstein, Chen, Smith, and Maldonado (2014) comment on persistence of women in engineering based on college demographics, by describing how women are well-represented in college, but still are poorly represented in engineering. Institutions are thus still struggling to recruit and retain women in engineering. Lichtenstein et al. (2014) attribute the high drop-out of women and other under-represented minorities from engineering due to a similar chilly unwelcome climate of the 4-year institutions. A potential follow up qualitative analysis could thus specifically be designed to understand the different experiences of students based on whether or not they are minorities in the discipline lead to differences in their career preparedness.

5.3 Limitations of this Research Study and Overall Research Quality

There are two limitations of this research study. These limitations do not invalidate the study, however, they do constrain the conclusions and the interpretations that can be drawn from these conclusions. The first set of limitations is related to the dataset analyzed. The dataset did not include raw input from the participants, but rather comprised excerpts that were isolated from the transcripts by an expert. These excerpts may have been (albeit, unintentionally) chosen to have similar word and phrase compositions and as a result there did not show up any significant differences in the word compositions and topic compositions for excerpts from male versus female, or excerpts from Juniors versus Seniors. That there exists a difference in the word choices, based on

the words ordered by tf-idf scores, indicates possibilities of different priorities in terms of factors contributing to career preparedness across the two groups of comparison. The second set of limitations is aligned with the limitations common to Natural Language Processing techniques such as the limited knowledge of semantics or context. Jurafsky and Martin (2007) describe semantics as the knowledge of meaning. Jurafsky and Martin note that language processing systems differ from other data processing systems since they use knowledge of the language. For instance, they identify a Unix program which may be used to count bytes and lines in a text file, as an ordinary data processing application. However, when the program is used to count the number of words in the file, the program requires the knowledge of "what it means to be a word" before it can begin counting. Training the machine to understand knowledge, contexts, and meaning is a non-trivial task, and require large sets of human annotated training data, which this study was limited in not having.

Keeping in mind the need to maintain a high quality for this research, I met regularly with the expert researchers during our team meetings to review my emergent results and discuss my findings. Montfort, et al. (2013) describe how meeting with expert researchers and asking questions about the content by the novice researcher, are important to maintain rigor and value of the method. Additionally, this findings of this research are presented with methodological transparency, detailing all the decisions made at every step of the process. This detailing may be helpful for Engineering Education researchers to understand ways in which the exploratory techniques were used. It is hoped that this research will lead to more researchers in Engineering Education appreciating the value

that Natural Language Processing techniques can offer to exploratory studies for qualitative datasets in the field.

Acknowledgement

This study is part of a parent study funded by the National Science Foundation under EEC-1360665, 1360956, and 1360958.

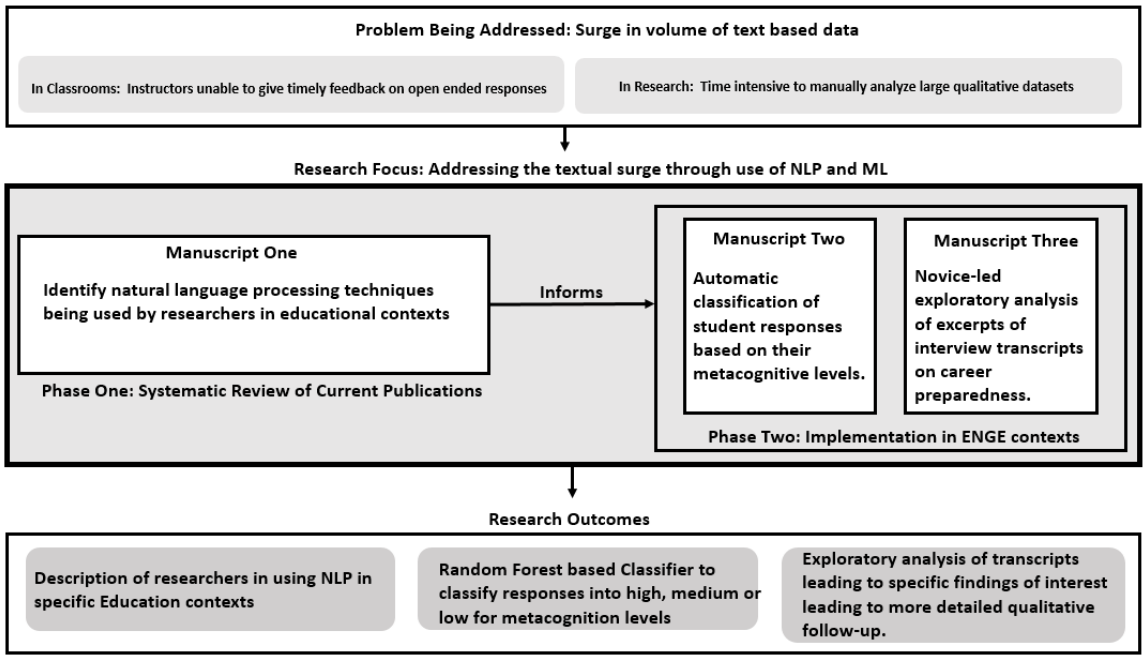
CHAPTER 5 DISCUSSION AND CONCLUSION

1. Introduction

The purpose of my doctoral research was two pronged: to describe the current state of use of Natural Language Processing as it applies to the broader field of Education, and to demonstrate the use of Natural Language Processing techniques for two Engineering Education specific contexts of instruction and research respectively. Thus, the preceding three chapters each addressed the following aims: (1) to identify Natural Language Processing techniques being used by researchers in education contexts, (2) to use Natural Language Processing techniques to automate the classification of student responses to open-ended prompts based on metacognition levels of the response, and (3) to use Natural Language Processing techniques to conduct a novice-lead exploratory analysis of excerpts from engineering student interviews related to career preparedness.

In phase 1, I systematically reviewed publications to identify machine learning and Natural Language Processing techniques used by researchers in education related contexts, with a focus on the specific contexts they were used in and the challenges faced by the researchers in using such techniques (Manuscript One). Results from phase 1 informed phase two which involved implementing Natural Language Processing and machine learning techniques for (a) developing an automated classifier to classify student responses to open ended reflective prompts based on the metacognitive level as demonstrated through the response (Manuscript 2) and (b) conducting an exploratory

textual analysis using machine learning and Natural Language Processing techniques for a novice led analysis of excerpts from interview transcripts of undergraduate engineering students being asked about career preparedness (Manuscript 3). A mapping of the aims and outcomes for the three manuscripts, as they relate to each other and to the problem being addressed is shown in Figure 1 below.



Ch 5. Fig. 1 Dissertation Research Focus with overview of Problem Being Addressed and Outcomes

2. Contributions

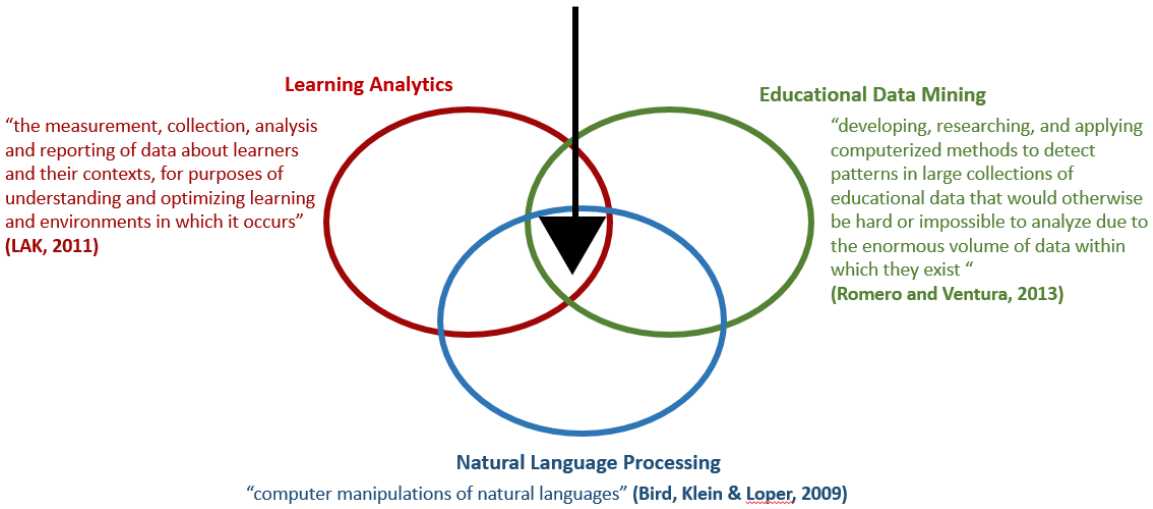
This research makes significant contributions which may be relevant for researchers not only in Engineering Education, but also in the broader fields of Natural Language Processing and Educational Data Mining. The specific contributions of this research are to: (1) expand on research at the intersection of contemporary research areas, (2) address

the call for new methodologies for emerging challenges in Engineering Education, (3) demonstrate the applicability of an innovative methodology to analyze qualitative datasets through statistical methods beyond counting, and (4) provide significant insights of relevance to multiple stake-holders.

2.1 Expand on research at the intersection of contemporary research areas

This research is a contribution to advancement of contemporary research areas of Learning Analytics, Natural Language Processing, and Education Data Mining. I have described in earlier sections how due to the increasing digitalization of data, educators, administrators, and researchers now have access to large amounts of student data. This has led to creation and advances in the fields such as Learning Analytics and Educational Data Mining. Learning Analytics may be understood by the definition provided by the 1st Learning Analytics and Knowledge Conference in 2011 as: “the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and environments in which it occurs.” (LAK, 2011). Related to learning analytics is the concept of Educational Data Mining, which Romero and Ventura (2013) described as: “developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist”.

My research, as depicted in Fig. 2 below, lies at the intersection of these contemporary research areas, infusing a Natural Language Processing lens to analyze student data, and mine information from it. All of these three research areas (i.e., Natural Language Processing, Learning Analytics and Educational Data Mining) are contemporary and research in these fields are upcoming. For example, Learning Analytics itself as a term gained popularity in 2011, with the first Learning Analytics and Knowledge conference which has now become an annual conference for educators researching in this area to present their work. Similarly, work on Educational Data Mining has garnered traction this past decade with the emergence and successful growth of several journals (such as the Journal of Educational Data Mining, started in 2009 and currently published its 9^{th} issue) devoted to research specifically in this area. My research, through presenting applicability of the concepts and methodologies to Engineering Education, presents how the scope of these contemporary and continually growing fields, may be extended to engineering classrooms.



Ch 5. Fig. 2 Research at the intersections of contemporary research areas

2.2 Address the call for new methodologies for emerging challenges in Engineering Education

This research is a contribution to Engineering Education by addressing the call for new methodologies for emerging challenges in the field. Papamitsiou and Econimides (2014) described how handling large amount of data manually is prohibitive, and detailed the growing need for sophisticated analytical techniques for educational contexts. Closer home, in Engineering Education literature, Case and Light (2011) urged the field to expand their methodological range in order to be able to answer diverse research questions targeting emerging challenges. I have described earlier how it is becoming increasingly important for instructors and researchers to be able to effectively and time

and resource efficiently channelize their productivity to analyzing student datasets. Using Natural Language Processing provides a statistical approach to increasing the efficiency, while at the same time providing research with a new methodology to analyze qualitative datasets.

2.3 Demonstrate applicability of an innovative methodology to analyze qualitative datasets using statistical methods beyond counting

This research is a direct contribution to expanding mixed methods research methodology by elaborating successful implementation of statistical methods to analyzing qualitative datasets beyond counting of keywords. Mixed Methods research typically combine a quantitative and a qualitative strand in analyzing datasets (Creswell, 2009). Often, the quantitative strand comprises numeric information derived from the dataset (e.g., demographic distribution of participants, age range of participants, etc.). In my research, I have used quantitative statistical techniques paired with a qualitative understanding of the words themselves to analyze textual datasets for two contexts. This research thus incorporates dialectical pluralism, which Greene and Hall (2010) described as a mode of inquiry that actively welcomes more than one mental model, methodology and method into the same inquiry space, engaging them in respectful dialogue with each other. This research elaborates on techniques and methods, which are inherently innovative and interdisciplinary, and may pave the future for interesting Mixed Methods research incorporating machine-based and statistically backed analysis of text data.

2.4 Provide significant insights of relevance to multiple stake-holders

This research contributes to the Engineering Education community by providing significant insights, which are of relevance to multiple stakeholders. For example, Soledad, et. al. (2017) described how open-ended responses on end-of-semester course-assessment surveys submitted by students are ignored for assessment due to analysis of these being time consuming. Through use of Natural Language Processing techniques such as the ones elaborated upon in this research, administrators may now be able to make meaning out of data such as these previously ignored open-ended responses to get deeper insights about the learning environments or student experiences. I have explained how another set of stakeholders: instructors may benefit from the insights of this research for direct implementation of such methodologies into their classroom or instruction. Similarly, for students, the insights of this research are significant for they may now be able to get faster and more personalized feedback, and can be ensured that their voices are heard. Finally, for novice and more experienced researchers, this research contributes by introducing a novel research methodology which is gaining more momentum in terms of techniques and which may become more widely researched and utilized in our field.

3. Implications

Results of my dissertation has implications related to both practice in engineering courses and in Engineering Education research. Implications for practice include providing engineering instructors an option to automatically and through a time-effective

process analyze student responses to open-ended prompts. Implications for research include the use of machine learning and natural language processing based techniques in different research studies, and also providing researchers who may use such tools a broad overview of existing research using such techniques including detailed analysis of limitations and contexts of such studies.

3.1 Implications for Practice

With increase in enrollment of students in the universities leading to larger classrooms, a challenge faced by instructors is that of providing timely and meaningful feedback to the students. Not only is it time consuming for individual instructors to grade responses manually, often it is also not possible to hire multiple graders for courses. Even when multiple instructors exist for the same class, or for small class sizes, a lot of time and resource is spent in assessing open-ended responses and providing feedback based on these or modifying instruction to accommodate changes based on the students' comments. However, reflective prompts may be helpful in gauging various skill development and impact of instruction on such competencies within the classrooms. Thus providing timely feedback is often a missed opportunity, especially in instances of open-ended reflective prompts which are often not graded or included in student assessments. Natural Language Processing techniques may help alleviate some of these problems faced by instructors in assessing open ended responses through use of automated classification, automated grading, or similar implementations. This could help potentially

free up instructors' time to then direct their resources and efforts to incorporate instructional changes to their instruction based on their understanding of the learning environment through results of the automatically graded responses.

My dissertation research has implications for practice through helping provide instructors with an alternative for dealing with assessing and providing feedback for student responses in their classrooms. In Manuscript One, I identified research studies that used Natural Language Processing for instructional tasks such as those related to automate grading. In my Manuscript One I have detailed the context for the implementations of these techniques, and have shown how a variety of Natural Language Processing techniques can be used for the different sets of tasks. I have also provided a discussion on strengths and weaknesses of use of such techniques as described in the current literature. My analysis shows that there are many researchers interested in implementing Natural Language Processing techniques to education contexts. However, the sources of publication of these results are not conveniently found through just searching in conventional education databases. Engineering educators will have to access multiple Computer Science databases such as ACM to get more access to studies using Natural Language Processing. In Manuscript Two, specifically for an Engineering Education context I elaborated how an automated classification system can be developed using Natural Language Processing to automate the classification of open ended student responses to High, Medium or Low, based on metacognitive levels described by experts. Systems like these may be of direct usefulness for instructors as some of them may struggle with assessing open-ended student responses due to the large classroom sizes.

3.2 Implications for Research

The research presented in my dissertation has implications for research through the use of Natural Language Processing techniques by engineering educators to analyze qualitative datasets. Using such techniques may be of interest to researchers, especially those working with large qualitative datasets, as well as to novice researchers working with experts to extract themes from large qualitative datasets.

Specifically, in Manuscript One I have provided an illustration of the different contexts that Machine Learning and Natural Language Processing have been used in for education research, mapping these contexts to context-specific challenges and limitations, and the more general limitations of using such methods. These results will thus equip Engineering Education researchers with a list of techniques being used, and an understanding of the challenges and limitations may help guide the study design for their research using such techniques. In Manuscript Three, I have detailed the use of multiple techniques (such as word clustering using Latent Dirichlet Allocation, tf-idf based term frequency comparisons, and Part of Speech specific analysis) for an Engineering Education dataset comprising excerpts from interview transcripts of engineering students elaborating about career preparedness. In Manuscript Three, apart from the methods I have also presented a list of challenges (such as the machine's limited knowledge of semantics, and the limitations due to the way the dataset was developed through excerpts chosen by an expert from interview transcripts) that were specific to my dataset, and led to limitations in my analysis. This knowledge may be transferable to other contexts as

researchers develop their study designs and think about analyzing data with similar characteristics through Natural Language Processing techniques. This research thus sets a path for further research in Engineering Education incorporating use of machine learning and natural language processing for exploratory analysis.

3.3 Summary of Implications

In summary, the results of my dissertation have important implications for both research and practice. These implications have been previously detailed in the individual manuscripts, and can be shown to advance the growing need and relevance of using machine learning and natural language processing techniques in Engineering Education instruction and research contexts.

4. Future Work and Challenges

While the findings of this study advance the relevance and scope of use of Natural Language Processing and machine learning in Engineering Education contexts, there are multiple avenues that can be explored in the realm of using automated textual analytics in our field. A direct next step of this research would be to address the prominent challenge which relates to the lack of an Engineering Education specific database. For example, the SNLI corpus, a collection of 570,000 human written English sentence pairs manually labeled for labels supporting the task of natural language inference (Bowman, et al. 2015). A labeled Engineering Education specific data base, similarly, may be helpful for

researchers in the field to use as basis for training their supervised machine learning based systems. We could also repeat the study in Manuscript Two to include a different group of students (e.g., more international students, or different level of education such as k-12 as opposed to undergraduate), and test the performance of the automated classifier. Since machine learning techniques are language agnostic, similar systems can also be tested in other languages, provided Part of Speech syntactical trainers are available for the language. Another scope for future research could be to develop a system which could convert conversations collected during an ongoing interview, to a visual diagram summarizing the topics of interest/focus, and the time spent on each broad topic, to help serve the interviewer assess the quality of the interview and the data immediately following the interview. This could also help the interviewer reflect upon the interview process and modify their interview techniques based on the research focus and thus collect data more effectively.

While Natural Language Processing use in social science research is increasingly fast paced, there are numerous challenges, which may impede this growth. Similar to the challenges identified for other fields such as finance, in Engineering Education, use and growth of Natural Language Processing based techniques may be impeded by the scarcity of data. While there is a recognized abundance of datasets in our field, labeled datasets are scarce. Without a sufficient number of labeled datasets, supervised machine learning is not possible, since we will not have adequate data to train and test machine learning based algorithms on. Labeling datasets are time consuming, however, researchers in the field have been qualitatively coding textual datasets, the codes for which may be used as

labels for the machine to learn and test on. A challenge in education research is that due to the nature of educational datasets and IRB regulations for distribution and use of these datasets, these labeled datasets are usually not made publicly available. As a result, unlike with analysis on other data sources such as sentiment analysis in Twitter or other social media sites where pre-labeled data from prior work in the field can easily be shared across research groups or assigned to human annotators through Mechanical Turks, education datasets are often restricted and more difficult to make openly available.

Another challenge is in terms of the large number of options available in terms of techniques themselves. In Manuscript One, I described how some authors elaborated on the challenges of using Natural Language Processing techniques due to the large number of pre-existing algorithms and their set of parameters. Learning about which algorithm is best fit for the particular dataset may be beyond the disciplinary expertise of many education researchers. As a result, researchers may often settle for data analytics packages, which use standard parameters optimized for different datasets, thus only yielding sub-par results for a particular dataset, which may discourage researchers from pursuing these methodologies.

Finally, the most challenging piece of incorporating machine intelligence in education research is that some of the algorithmic results are yet mathematically uninterpretable. Knight (2017) in their article in the Technology Review wrote about the "Dark Secret at the Heart of AI," and explained how machine learning and AI researchers have not yet been able to understand or demystify why certain systems are able to work correctly. Explaining this for a lay audience, Knight elaborates that since many of the AI's

decisions are made in multi-dimensional spaces which human beings are unable to visualize, the final decisions made are accurate yet remain unexplained mathematically, and thus may be thought of as intelligent black boxes. For example, Knight describes how the chip-maker Nvidia piloted an autonomous car which was not programmed by an engineer, but rather 'self-taught' itself to drive through access to human videos on driving. Knight describes how this is an impressive feat, but also unsettling since to researchers it is not completely clear how the car makes its decisions.

Transparency in the decisions made by autonomous systems are important in social science research, especially while dealing directly with humans. For example, Lipton (2016) in their paper titled "Mythos of Model Interpretability" elaborate how machine-based algorithmic, statistics-backed predictions may indeed be accurate and may have 'learned from existing data', the lack of interpretability makes the model and hence the results of the model subject to ambiguity in terms of trust, informativeness, causality, transferability, and ultimately making fair and ethical decisions. In effect, the author contributes to the ongoing debate regarding AI, on how we can make decisions based on uninterpretable outputs of machines. As an example the author cites the use of models to forecast crime rates for allocating police officers to areas. Lipton explains that although the model may be making accurate predictions, it may also not be accounting for racial bias in the training dataset, and thus may lead to a cycle of perpetuating over-policing in certain neighborhood. Lipton advises caution in relinquishing all control to machine outputs.

Thus, rightly, in fields such as education, results from black-box systems may face strong resistance from the community. This is already true for the field of economics, with the European Union from the Summer of 2018 requiring that companies be able to provide detailed explanations even if they are decisions that automated systems have reached for example those related to loan or mortgage approval or denial. Similarly, in Engineering Education, despite the advantages of incorporating algorithms in research and practice, there may be strong reluctance to rely on machine intelligence to make meaning out of data if the machine intelligence itself is difficult to understand or interpret. The onus lies on the early movers in this field to ensure that transparent methodologies are being implemented and adequate rationale is present at every stage. Thus, while the conversation around machine intelligence often leads to debate on machines replacing human experts, for education, incorporating machine intelligence would require higher level of knowledge and expertise from the humans to ensure that the systems are accurately and transparently being used to serve the community.

5. Concluding Remarks

Having served as an instructor in a large freshman engineering course, and as a researcher associated with the Office of Assessment and Evaluation at Virginia Tech, I have faced the challenges of manually assessing and analyzing qualitative student-related data. Qualitative data such as open-ended responses to reflective prompts are often not included in analysis, due to their analysis being time consuming and resource intensive. Natural Language Processing used in conjunction with machine learning may provide a

feasible tool for analyzing such qualitative datasets, and obtaining results and deeper insights from them. As universities move towards more electronic databases, using automated intelligence to extract information from textual datasets may be a sustainable solution for tackling this ever-increasing volume of qualitative data. I would like to see more Engineering Education instructors and researchers use these techniques, so that open-ended reflective responses no longer remain a missed opportunity and overviews for large qualitative datasets may be formed quickly. Ultimately, through my dissertation research I hope to have demonstrated how an interdisciplinary approach using computational technologies and Engineering Education related knowledge can be used to solve the problem of analyzing large textual data to contribute to further development of the field.

References

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics Learning analytics (pp. 61-75): Springer.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391-402.

Becker, L., Palmer, M., van Vuuren, S., & Ward, W. (2012, June). Question ranking and selection in tutorial dialogues. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 1-11). Association for Computational Linguistics.

Bethard, S., Hang, H., Okoye, I., Martin, J. H., Sultan, M. A., & Sumner, T. (2012, June). Identifying science concepts and student misconceptions in an interactive essay writing tutor. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 12-21). Association for Computational Linguistics.

Bhaduri, S. (in preperation). NLP in the ENGE classrooms: using automatic text classification in gauging metacognitive levels.

Bhaduri, S. (in preperation). Systematically Exploring the Use of Natural Language Processing in Education.

Bhaduri, S. and T. Roy (2017, June). Demonstrating Use of Natural Language Processing to Compare College of Engineering Mission Statements. 2017 *ASEE Annual Conference*. Columbus, Ohio.

Bhaduri, S., & Roy, T. (2017, October). A word-space visualization approach to study college of engineering mission statements. In *2017 IEEE Frontiers in Education Conference (FIE)* (pp. 1-5). IEEE.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit: " O'Reilly Media, Inc.".

Blair, E., & Noel, K. V. (2014). Improving higher education practice through student evaluation systems: is the student voice being heard? Assessment & Evaluation in Higher Education, 39(7), 879-894.

Board, N. S. (2016). Science & engineering indicators (Vol. 1): National Science Board.

Borrego, M., Foster, M. J., & Froyd, J. E. (2014). Systematic literature reviews in engineering education and other developing interdisciplinary fields. Journal of Engineering Education, 103(1), 45-76.

Borrego, M., Foster, M. J., & Froyd, J. E. (2015). What Is the State of theArt of Systematic Reviewin Engineering Education? Journal of Engineering Education, 104(2), 212-242. doi:10.1002/jee.20069

Bravo-Marquez, F., L'Huillier, G., Moya, P., Ríos, S. A., & Velásquez, J. D. (2011, September). An automatic text comprehension classifier based on mental models

and latent semantic features. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* (p. 23). ACM.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brown, A. L. (1997). Transforming schools into communities of thinking and learning about serious matters. American psychologist, 52(4), 399.

Brown, P. R., McCord, R. E., Matusovich, H. M., & Kajfez, R. L. (2015). The use of motivation theory in engineering education research: a systematic review of literature. European journal of engineering education, 40(2), 186-205.

Calvo, R. A., & Ellis, R. A. (2010). Students' conceptions of tutor and automated feedback in professional writing. *Journal of Engineering Education*, 99(4), 427-438.

Carrico, C., Harris, A., Matusovich, H. M., Brunhaver, S. R., Streveler, R. A., & Sheppard, S. (2016, June). Helping engineering students get jobs: Views from career services professionals. In 123rd ASEE Annual Conference and Exposition. American Society for Engineering Education.

Chakraborty, U. K., Konar, D., Roy, S., & Choudhury, S. (2016). Intelligent fuzzy spelling evaluator for e-Learning systems. *Education and Information Technologies*, *21*(1), 171-184.

Chen, E. (2011, April 27). Choosing a Machine Learning Classifier. Retrieved January 01, 2018, from http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/

Cormack, G. V. (2008). Email spam filtering: A systematic review. Foundations and Trends® in Information Retrieval, 1(4), 335-455.

Creamer, E. G., Simmons, D., & Yu, R. (to be published). Procedures for Conducting a Systematic Mixed Methods Literature Synthesis

Creswell, J. W. (2014). Research design: qualitative, quantitative, and mixed method approaches (4th ed.). Thousand Oaks, Calif: Sage Publications.

Creswell, J. W., & Plano Clark, V. L. (2007). Designing and conducting mixed methods research. Thousand Oaks, Calif: SAGE Publications.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. Bethesda (Maryland): National Institutes of Health, 2094-2103.

Cunningham, P., Matusovich, H. M., Hunter, D. A., & McCord, R. E. (2015). Teaching metacognition: Helping engineering students take ownership of their own learning. Paper presented at the Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE.

Cunningham, P., Matusovich, H., Morelock, J., & Hunter, D.-A. N. (2016). Beginning to Understand and Promote Engineering Students' Metacognitive Development. Paper presented at the American Society for Engineering Education.

Cunningham, P., Williams, S., Matusovich, H., & Bhaduri, S. (2017). Beginning to Understand Student Indicators of Metacognition. Paper presented at the American Society for Engineering Education.

Cuseo, J. (2007). The empirical case against large class size: adverse effects on the teaching, learning, and retention of first-year students. The Journal of Faculty Development, 21(1), 5-21.

Cutrone, L. A., & Chang, M. (2010, July). Automarking: automatic assessment of open questions. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on* (pp. 143-147). IEEE.

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., Bianco, M., & McNamara, D. S. (2015, March). ReaderBench: An integrated tool supporting both individual and collaborative learning. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 436-437). ACM.

Davies, P. (2000). The Relevance of Systematic Reviews to Educational Policy and Practice. Oxford Review of Education, 26(3/4), 365-378.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? Journal of biomedical informatics, 42(5), 760-772.

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, *24*(3), 284-332.

Echeverría, V., Gomez, J. C., & Moens, M. F. (2013, December). Automatic labeling of forums using bloom's taxonomy. In *International Conference on Advanced Data Mining and Applications* (pp. 517-528). Springer, Berlin, Heidelberg.

Evidence for Policy and practice Information and Co-ordinating Centre (EPPI-Centre). (2010). EPPI-Centre methods for conducting systematic reviews. Retrieved from London:

Feng, H. H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2016). Automated error detection for developing grammar proficiency of ESL learners. *calico journal*, *33*(1), 49.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American psychologist, 34(10), 906.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906-914.

Gomaa, W. H., & Fahmy, A. A. (2012). Short answer grading using string similarity and corpus-based similarity. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *3*(11).

Gough, D., Oliver, S., & Thomas, J. (2012). Introducing Systematic Reviews. In D. Gough, S. Oliver, & J. Thomas (Eds.), An introduction to systematic reviews: Sage.

Greene, J. C. (2008). Is Mixed Methods Social Inquiry a Distinctive Methodology? Journal of Mixed Methods Research, 2(1), 7-22. doi:10.1177/1558689807309969

Heilman, M., & Smith, N. A. (2010, June). Extracting simplified statements for factual question generation. In *Proceedings of QG2010: The Third Workshop on Ques-tion Generation* (p. 11).

Heyvaert, M., Maes, B., & Onghena, P. (2013). Mixed methods research synthesis: definition, framework, and potential. Quality & Quantity, 47(2), 659-676.

Hiles, D. R. (2008). Transparency. The Sage Encyclopedia of Qualitative Research Methods. SAGE Publications, Inc (pp. 891-893). Thousand Oaks, CA: SAGE Publications, Inc.

Hoshino, A., & Nakagawa, H. (2005, October). WebExperimenter for multiple-choice question generation. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* (pp. 18-19). Association for Computational Linguistics.

Jayakodi, K., Bandara, M., Perera, I., & Meedeniya, D. (2016). WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom's Taxonomy. *International Journal of Emerging Technologies in Learning*, 11(4).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, 137-142.

Jurafsky, D., & Martin, J. H. (2007). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Vol. 26).

Knight, W. (2017, April 11). The Dark Secret at the Heart of AI. Retrieved January 01, 2018, from https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Lan, A. S., Vats, D., Waters, A. E., & Baraniuk, R. G. (2015, March). Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 167-176). ACM.

Leeman-Munk, S. P., Wiebe, E. N., & Lester, J. C. (2014). Assessing elementary students' science competency with text analytics. Paper presented at the Proceedings of the Fourth International Conference on Learning Analytics And Knowledge.

Leydens, J. A., Moskal, B. M., & Pavelich, M. (2004). Qualitative methods used in the assessment of engineering education. Journal of Engineering Education, 93(1), 65-72.

Lichtenstein, G., Chen, H. L., Smith, K. A., & Maldonado, T. A. (2014). Retention and persistence of women and minorities along the engineering pathway in the United States. Handbook of engineering education research, 311-334.

Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.

Liu, C. L., Lin, J. H., & Wang, Y. C. (2010). Applications of NLP Techniques to Computer-Assisted Authoring of Test Items for Elementary Chinese. *Online Submission*, *7*(3), 42-54.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. Behavior Research Methods, 44(3), 608-621.

Maskeri, G., Sarkar, S., & Heafield, K. (2008, February). Mining business topics in source code using latent dirichlet allocation. In *Proceedings of the 1st India software engineering conference* (pp. 113-120). ACM.

McCord, R., & Matusovich, H. M. (2013). Developing an Instrument to Measure Motivation, Learning Strategies and Conceptual Change. Paper presented at the 120th ASEE Annual Conference & Exposition, Atlanta, GA.

McGowan, J., & Sampson, M. (2005). Systematic reviews need systematic searchers. Journal of the Medical Library Association, 93(1), 74.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011, June). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 752-762). Association for Computational Linguistics.

Montfort, D. B., Herman, G. L., Brown, S., Matusovich, H., & Streveler, R. (2013). Novice-led paired thematic analysis: A method for conceptual change in engineering. age, 23, 1.

Mora, H., Ferrández, A., Gil, D., & Peral, J. (2017). A Computational Method for Enabling Teaching-Learning Process in Huge Online Courses and Communities. *The International Review of Research in Open and Distributed Learning*, 18(1).

Morelock, J. R. (2017). A systematic literature review of engineering identity: definitions, factors, and interventions affecting development, and means of measurement. European journal of engineering education, 1-23.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006, May). Analyzing entities and topics in news articles using statistical topic models. In *ISI* (pp. 93-104).

Oblinger, D. G. (2012). Let's Talk... Analytics. EducausE Review, 47(4), 10-13.

O'Cathain, A., Murphy, E., & Nicholl, J. (2008). The quality of mixed methods studies in health services research. Journal of Health Services Research & Policy, 13(2), 92-98. doi:10.1258/jhsrp.2007.007074

Ozturk, Z. K., Cicek, Z. E., & Ergul, Z. (2017). Sentiment Analysis: an Application to Anadolu University. Acta Physica Polonica A, 132(3), 753-755.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49.

Parry, M. (2012). " Supersizing" the College Classroom: How One Instructor Teaches 2,670 Students. Chronicle of Higher Education.

Patton, M. Q. (2005). Qualitative research: Wiley Online Library.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP) (pp. 1532-1543).

Persing, I., Davis, A., & Ng, V. (2010, October). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 229-239). Association for Computational Linguistics.

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide Note Classification Using Natural Language Processing: A Content Analysis. Biomedical Informatics Insights, 2010(3), 19-28. doi:10.4137/BII.S4706

Petticrew, M., & Roberts, H. (2008). Systematic reviews in the social sciences: A practical guide: John Wiley & Sons.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. Theory into practice, 41(4), 219-225.

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education*, 1-48.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016). Forecasting student achievement in MOOCs with natural language processing. Paper presented at the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Sandelowski, M. (2008). Reading, writing and systematic review. Journal of advanced nursing, 64(1), 104-110. doi:10.1111/j.1365-2648.2008.04813.x

Sandelowski, M., Voils, C. I., & Barroso, J. (2006). Defining and Designing Mixed Research Synthesis Studies. Research in the schools : a nationally refereed journal sponsored by the Mid-South Educational Research Association and the University of Alabama, 13(1), 29.

Sandelowski, M., Voils, C. I., Barroso, J., & Lee, E. J. (2008). “Distorted into clarity”: A methodological case study illustrating the paradox of systematic review. Research in nursing & health, 31(5), 454-465. doi:10.1002/nur.20278

Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., & Chute, C. G. (2008). Mayo clinic NLP system for patient smoking status identification. Journal of the American Medical Informatics Association, 15(1), 25-28.

Savoy, J. (2015). Text clustering: An application with the State of the Union addresses. Journal of the Association for Information Science and Technology, 66(8), 1645-1654.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Quality & Quantity, 47(2), 761-773. doi:10.1007/s11135-011-9545-7

Scheponik, T., Sherman, A. T., DeLatte, D., Phatak, D., Oliva, L., Thompson, J., & Herman, G. L. (2016, October). How students reason about Cybersecurity concepts. In *Frontiers in Education Conference (FIE), 2016 IEEE* (pp. 1-5). IEEE.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34(1), 1-47. doi:10.1145/505282.505283

Shaw, E. (2014). PLAY Minecraft! Assessing secondary engineering education using game challenges within a participatory learning environment. In *Proceedings of the 121 Annual Conference and Exposition, 360 of Engineering Education, paper 8438, Indianapolis, Indiana*.

Shukla, H., & Kakkar, M. (2016, January). Keyword extraction from Educational Video transcripts using NLP techniques. *In Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference* (pp. 105-108). IEEE.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EducausE Review, 46(5), 30.

Sil, A., Ketelhut, D. J., Shelton, A., & Yates, A. (2012, June). Automatic grading of scientific inquiry. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 22-32). Association for Computational Linguistics.

Skinner, J. K. (2015). Bibliometric and social network analysis of doctoral research: Research trends in distance learning (Doctoral dissertation, The University of New Mexico).

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57(10), 1510-1529.

Soledad, M., & Grohs, J. (2017). Understanding Faculty Experiences in Teaching Large Classes: A Pilot Study on Faculty Perceptions of Teacher-Student Interaction in Foundational Engineering Courses.

Soledad, M., Grohs, J., Bhaduri, S., Doggett, J., Williams, J., & Culver, S. (2017). Leveraging Institutional Data to Understand Student Perceptions of Teachingin Large Engineering Classes. Paper presented at the 2017 IEEE Frontiers in Education Conference (FIE) (FIE 2017), Indianapolis, USA.

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. Journal of the American Medical Informatics Association, 17(6), 646-651.

Sverdlik, Y. (2017, September 29). Financial Sector's Top Five AI Challenges, According to Morgan Stanley. Retrieved January 01, 2018, from http://www.datacenterknowledge.com/machine-learning/financial-sectors-top-five-ai-challenges-according-morgan-stanley

Teddlie, C., & Tashakkori, A. (2012). Common "core" characteristics of mixed methods research A review of critical issues and call for greater convergence. American Behavioral Scientist, 56(6), 774-788.

Tonso, K. L. (2006). Teams that work: Campus culture, engineer identity and social interactions. Journal of engineering education (Washington, D.C.), 95, 25-37.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British journal of management, 14(3), 207-222.

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2014). PolyCAFe—automatic support for the polyphonic analysis of CSCL chats. *International Journal of Computer-Supported Collaborative Learning*, *9*(2), 127-156.

Variawa, C. (2014). Investigating the Language of Engineering Education. University of Toronto.

Variawa, C., McCahan, S., & Chignell, M. (2013). An Automated Approach for Finding Course-specific Vocabulary. Paper presented at the 2013 ASEE Annual Conference & Exposition, Atlanta, Georgia.

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. Metacognition and learning, 1(1), 3-14.

Voils, C. I., Sandelowski, M., Barroso, J., & Hasselblad, V. (2008). Making sense of qualitative and quantitative findings in mixed research synthesis studies. Field methods, 20(1), 3-25.

Watanabe, W. M., Candido Jr, A., Amâncio, M. A., De Oliveira, M., Pardo, T. A. S., Fortes, R. P., & Aluísio, S. M. (2010). Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. *New Review of Hypermedia and Multimedia, 16*(3), 303-327.

Wen, M., Yang, D., & Rose, C. (2014, July). Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In *Educational data mining 2014*.

Xian, H., & Madhavan, K. (2014). Anatomy of Scholarly Collaboration in Engineering Education: A Big-Data Bibliometric Analysis. *Journal of Engineering Education, 103*(3), 486-514.

Xiong, W., & Litman, D. (2011, June). Understanding differences in perceived peer-review helpfulness using natural language processing. In *Proceedings of the 6th*

Workshop on Innovative Use of NLP for Building Educational Applications(pp. 10-19). Association for Computational Linguistics.

Xiong, W., Litman, D. J., & Schunn, C. D. (2010). Assessing reviewers performance based on mining problem localization in peer-review data.

Xiong, W., Litman, D., Wang, J., & Schunn, C. (2012, June). An interactive analytic tool for peer-review exploration. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 174-179). Association for Computational Linguistics.

Yannakoudakis, H., & Briscoe, T. (2012, June). Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 33-43). Association for Computational Linguistics.

Zeng-Treitler, Q., Perri, S., Nakamura, C., Kuang, J., Hill, B., Bui, D. D. A., . . . Bray, B. E. (2014). Evaluation of a pictograph enhancement system for patient instruction: a recall study. Journal of the American Medical Informatics Association, 21(6), 1026-1031.

Zhu, Y., Yan, E., & Song, I.-Y. (2016). A natural language interface to a graph-based bibliographic information retrieval system. Sandelowski, M., Leeman, J., Knafl, K., & Crandell, J. L. (2013). Text-in-context: a method for extracting findings in mixed-methods mixed research synthesis studies. Journal of advanced nursing, 69(6), 1428-1437.

Appendix One: List of Articles Included in Systematic Review

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391-402.

Becker, L., Palmer, M., van Vuuren, S., & Ward, W. (2012, June). Question ranking and selection in tutorial dialogues. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 1-11). Association for Computational Linguistics.

Bethard, S., Hang, H., Okoye, I., Martin, J. H., Sultan, M. A., & Sumner, T. (2012, June). Identifying science concepts and student misconceptions in an interactive essay writing tutor. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 12-21). Association for Computational Linguistics.

Bhaduri, S. and T. Roy (2017). Demonstrating Use of Natural Language Processing to Compare College of Engineering Mission Statements. ASEE Annual Conference Proceedings. Columbus, Ohio.

Bravo-Marquez, F., L'Huillier, G., Moya, P., Ríos, S. A., & Velásquez, J. D. (2011, September). An automatic text comprehension classifier based on mental models and latent semantic features. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* (p. 23). ACM.

Calvo, R. A., & Ellis, R. A. (2010). Students' conceptions of tutor and automated feedback in professional writing. *Journal of Engineering Education*, 99(4), 427-438.

Chakraborty, U. K., Konar, D., Roy, S., & Choudhury, S. (2016). Intelligent fuzzy spelling evaluator for e-Learning systems. *Education and Information Technologies*, *21*(1), 171-184.

Cutrone, L. A., & Chang, M. (2010, July). Automarking: automatic assessment of open questions. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on* (pp. 143-147). IEEE.

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., Bianco, M., & McNamara, D. S. (2015, March). ReaderBench: An integrated tool supporting both individual and collaborative learning. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 436-437). ACM.

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, *24*(3), 284-332.

Echeverría, V., Gomez, J. C., & Moens, M. F. (2013, December). Automatic labeling of forums using bloom’s taxonomy. In *International Conference on Advanced Data Mining and Applications* (pp. 517-528). Springer, Berlin, Heidelberg.

Feng, H. H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2016). Automated error detection for developing grammar proficiency of ESL learners. *calico journal*, *33*(1), 49.

Gomaa, W. H., & Fahmy, A. A. (2012). Short answer grading using string similarity and corpus-based similarity. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *3*(11).

Heilman, M., & Smith, N. A. (2010, June). Extracting simplified statements for factual question generation. In *Proceedings of QG2010: The Third Workshop on Ques-tion Generation* (p. 11).

Jayakodi, K., Bandara, M., Perera, I., & Meedeniya, D. (2016). WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom's Taxonomy. *International Journal of Emerging Technologies in Learning*, 11(4).

Lan, A. S., Vats, D., Waters, A. E., & Baraniuk, R. G. (2015, March). Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 167-176). ACM.

Liu, C. L., Lin, J. H., & Wang, Y. C. (2010). Applications of NLP Techniques to Computer-Assisted Authoring of Test Items for Elementary Chinese. *Online Submission*, *7*(3), 42-54.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011, June). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 752-762). Association for Computational Linguistics.

Mora, H., Ferrández, A., Gil, D., & Peral, J. (2017). A Computational Method for Enabling Teaching-Learning Process in Huge Online Courses and Communities. *The International Review of Research in Open and Distributed Learning*, 18(1).

Ozturk, Z. K., Cicek, Z. E., & Ergul, Z. (2017). Sentiment Analysis: an Application to Anadolu University. Acta Physica Polonica A, 132(3), 753-755.

Persing, I., Davis, A., & Ng, V. (2010, October). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 229-239). Association for Computational Linguistics.

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education*, 1-48.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016, April). Forecasting student achievement in MOOCs with natural language processing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 383-387). ACM.

Shaw, E. (2014). PLAY Minecraft! Assessing secondary engineering education using game challenges within a participatory learning environment. In *Proceedings of the 121 Annual Conference and Exposition, 360 of Engineering Education, paper 8438, Indianapolis, Indiana*.

Sil, A., Ketelhut, D. J., Shelton, A., & Yates, A. (2012, June). Automatic grading of scientific inquiry. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 22-32). Association for Computational Linguistics.

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2014). PolyCAFe—automatic support for the polyphonic analysis of CSCL chats. *International Journal of Computer-Supported Collaborative Learning*, *9*(2), 127-156.

Variawa, C., et al. (2013). An Automated Approach for Finding Course-specific Vocabulary. 2013 ASEE Annual Conference & Exposition, Atlanta, Georgia.

Watanabe, W. M., Candido Jr, A., Amâncio, M. A., De Oliveira, M., Pardo, T. A. S., Fortes, R. P., & Aluísio, S. M. (2010). Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. *New Review of Hypermedia and Multimedia*, *16*(3), 303-327.

Wen, M., Yang, D., & Rose, C. (2014, July). Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In *Educational data mining 2014*.

Xian, H., & Madhavan, K. (2014). Anatomy of Scholarly Collaboration in Engineering Education: A Big-Data Bibliometric Analysis. *Journal of Engineering Education*, *103*(3), 486-514.

Xiong, W., & Litman, D. (2011, June). Understanding differences in perceived peer-review helpfulness using natural language processing. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*(pp. 10-19). Association for Computational Linguistics.

Xiong, W., Litman, D. J., & Schunn, C. D. (2010). Assessing reviewers performance based on mining problem localization in peer-review data.

Xiong, W., Litman, D., Wang, J., & Schunn, C. (2012, June). An interactive analytic tool for peer-review exploration. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 174-179). Association for Computational Linguistics.

Yannakoudakis, H., & Briscoe, T. (2012, June). Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 33-43). Association for Computational Linguistics.