Data-Science Perceptions: A Textual Analysis of Reddit Posts from Non-Computing Engineers

Mr. Nicolas Leger, Florida International University

Nicolas Léger is currently an engineering and computing education Ph.D. student in the School of Universal Computing, Construction, and Engineering Education (SUCCEED) at Florida International University. He earned a B.S. in Chemical and Biomolecular Engineering from the University of Maryland at College Park in May 2021 and began his Ph.D. studies the following fall semester. His research interests center on numerical and computational methods in STEM education and in Engineering Entrepreneurship.

Maimuna Begum Kali, Florida International University

Maimuna Begum Kali is a Ph.D. candidate in the Engineering and Computing Education program at the School of Universal Computing, Construction, and Engineering Education (SUCCEED) at Florida International University (FIU). She earned her B.Sc. in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET). Kali's research interests center on exploring the experiences of marginalized engineering students, with a particular focus on their hidden identity, mental health, and wellbeing. Her work aims to enhance inclusivity and diversity in engineering education, contributing to the larger body of research in the field.

Stephanie Jill Lunn, Florida International University

Stephanie Lunn is an Assistant Professor in the School of Universal Computing, Construction, and Engineering Education (SUCCEED) and the STEM Transformation Institute at Florida International University (FIU). She also has a secondary appointment in the Knight Foundation School of Computing and Information Sciences (KFSCIS). Previously, Dr. Lunn earned her doctorate in computer science from the KFSCIS at FIU, with a focus on computing education. She also holds B.S. and M.S. degrees in computer science from FIU and B.S. and M.S. degrees in neuroscience from the University of Miami. In addition, she served as a postdoctoral fellow in the Wallace H. Coulter Department of Biomedical Engineering at the Georgia Institute of Technology, with a focus on engineering education. Her research interests span the fields of computing and engineering education, human-computer interaction, data science, and machine learning.

Data Science Perceptions: A Textual Analysis of Reddit Posts from Non-Computing Engineers

Abstract

National reports in the United States regularly emphasize the need for qualified engineers to enter the workforce to solve present and future challenges for society. Such advancements often encourage an understanding and application of data science, a field that combines areas like mathematics, statistics, programming, analytics, and artificial intelligence. Despite its rapid growth and increasing integration across topics and industries, data science is not often incorporated directly into engineering curricula. Understanding when and how to utilize data science methodologies can provide non-computing engineers with a competitive edge professionally, offering valuable insights, improving decision-making, and driving innovation in their respective domains. Given the benefits of learning and employing data science, we explored the views of non-computing engineers and how they may influence their attitudes and practices. We defined non-computing engineers as individuals focused on an engineering field who are not pursuing computer science or computer engineering-specific formal education or degrees. To assess varying perspectives, we conducted a study utilizing Reddit posts. Reddit is a platform where many engineering students and practitioners may talk openly about different topics. We collected data using web scraping and analyzed it using a couple of Natural Language Processing (NLP) techniques, including Latent Dirichlet Allocation (LDA). Using the top keywords, we then took a manual approach, using whole posts for context to perform thematic analysis to derive the topics. Our findings suggest that non-computing engineers are generally positive about data science and its potential applications. They see it as especially important for 1) Career Prospects and Opportunities; 2) Ongoing Professional Development and Upskilling; and 3) Practical Applications. As such, it can provide opportunities for career preparedness, fostering new competencies, and a need to gain hands-on experience using data science to create value and solve problems. The results of this work can have important implications for educators, administrators, and professionals looking to incorporate data science into engineering praxis.

Keywords: Data Science, Non-Computing Engineers, Technology Acceptance Model, Reddit, LDA, Web Scraping

1. Introduction

Data science is an interdisciplinary field that involves extracting knowledge and insights from data (i.e., a collection of information or facts) using scientific methods, algorithms, and tools [1]. It encompasses a wide range of techniques, including statistical analysis, machine learning, data mining, and data visualization. The primary goal of data science is to uncover patterns, trends, and correlations within data to drive informed decision-making and solve complex problems. In various domains, such as healthcare [2], finance [3], marketing [4], and social media [5], data science has become increasingly important. It has transformed these fields by enabling the collection, storage, and processing of large volumes of data, which were previously impractical or impossible to handle.

With the application of data science techniques, valuable insights can be derived from these datasets, leading to improved strategies, predictions, and overall performance. For instance, in the realm of social media, data science has brought about a paradigm shift in the understanding of communication. It has moved beyond analyzing communication as signs or discourse and now encompasses the collection, storage, and processing of communication data. This expansion in perspective has opened up new possibilities for studying and leveraging social media platforms in various domains. For example, at the earlier stage of social media, Langlois et al. proposed an ontological shift, suggesting that with the help of data science, "we must expand from the study of communication as signs or discourse to include the study of communication as data collection, storage, and processing [5, p. 2]." Consequently, these new technologies further expand some fields as we know them.

There is also a growing body of work looking at data science applications in engineering [6]. Although we know it may be applied or beneficial for the broader field and its subfields (e.g., mechanical, industrial, chemical), we are limited in our understanding of how non-computing engineers may apply it in their work or practice. With that said, it is necessary to understand how non-computing engineers may apply data science in their work, as this remains a challenge in the field. In the context of engineering education and practice, Beck et al.'s article suggests adding data science as a "competency" in chemical engineering both in "the university curriculum or in a professional development context." They also give some clear examples of how data science principles can be used and potentially implemented in the traditional "numerical methods or applied computing course [7, p. 1413]." In this work, we define non-computing engineers as those individuals who are not pursuing computer science or computer engineering-specific formal education or degrees.

However, despite the rapid growth and increasing adoption of data science in industry, there remain challenges in incorporating it into engineering learning settings that traditionally have not heavily utilized data science computing techniques. These issues arise from the interdisciplinary character of data science, as educators struggle to integrate the application of professional

knowledge to real-world teaching scenarios while learners struggle to master each of the component fields [8]. Non-computing engineers can benefit from understanding how to apply data science methodologies in their work, as it can provide them with valuable insights, improve decision-making, and drive innovation in their respective domains. As technology progresses, data science practices [9] and computer-based tools [10] continue to expand and advance, they have become increasingly integrated into the engineering world. According to a study conducted by the McKinsey Global Institute (MGI) in 2011, the analysis of massive data sets will become crucial to competitiveness, productivity development, and innovation. They note that "in manufacturing, integrating data from Research and Development (R&D), engineering, and manufacturing units to enable concurrent engineering can significantly cut time to market and improve quality [11, p. 5]." While the need for data science may be clear, the extent to which it is used and applied in other fields like engineering is less well defined. Consequently, the research question (RQ) guiding this work is:

What is the perceived usefulness of data science as described by non-computing engineers?

In order to answer this question, we identified the Reddit platform as a tool to delve further into informal perceptions of data science, given the range of possible responses on different engineering topics, and its accessibility by the general public. Given that Reddit is a public forum, the user type is not discernible, allowing for a combination of those who may engage with engineering at different stages of their personal and/or professional life cycle. As such, we use the general term "Reddit users" when considering any potential person who may have posted in the forums of interest (more explanation is provided in the Methods). Toward our goal, the data was collected using web scraping and analyzed using Natural Language Processing (NLP) techniques; more information is provided in the sections below (see Section 4).

The paper is organized in the following way: In Section 2, we will discuss the theoretical framework that guided our work. After providing backgrounds in Section 3, we then overview our approach in Section 4. Section 5 presents the findings, and we wrap up the paper with a discussion (Section 6), limitations (Section 7), and conclusions (Section 8).

2. Theoretical Framework

In this paper, we applied the Technology Acceptance Model (TAM) as the guiding theoretical framework. TAM is a theoretical framework in the field of information systems research and has been widely used to study individuals' acceptance and usage behavior toward new technology [12]. TAM proposes that individuals' acceptance and usage behavior of technology are primarily determined by two key factors: perceived usefulness (U) and perceived ease of use (E) (*See Fig.1*) [12]. In this model, perceived usefulness refers to the extent to which an individual believes that using a particular technology will enhance his or her performance or productivity, while perceived ease of use refers to the extent to which an individual believes that using a

particular technology will be easy and effortless [13]. It has been demonstrated that both (U) and (E) are important predictors of people's intentions to use technology, which eventually results in real usage behavior [14], [15].

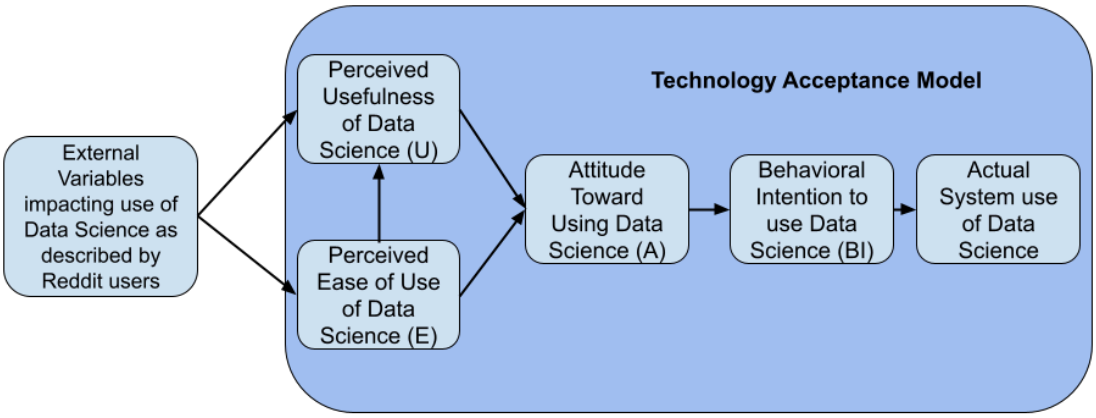

Figure 1. Adapted Technology Acceptance Model (TAM) [15]

TAM also incorporates external variables that may influence individuals' attitudes and behavior toward technology, such as social influence and facilitating conditions [13]. **Social influence** refers to the extent to which an individual's behavior is influenced by the opinions of others, such as peers and colleagues while facilitating conditions refer to the extent to which an individual believes that he or she has the necessary resources and support to use technology effectively. TAM has been widely used to study various technologies and contexts, such as e-commerce [16], e-learning tools adoption [17], and healthcare technologies [18]. Overall, TAM provides a valuable framework for understanding individuals' acceptance and usage behavior toward new technology.

For this paper, in the context of engineering problem-solving, perceived usefulness refers to the extent to which data science tools and techniques are perceived to improve the quality and efficiency of problem-solving, while perceived ease of use refers to the extent to which these tools and techniques are perceived to be easy to learn and applied. By incorporating the key factors of perceived usefulness and perceived ease of use and external variables, TAM enables researchers and practitioners to gain insights into the complex dynamics of technology adoption and usage. We enlist this framework as the foundation for our investigation to learn more about how data science is used and considered when approaching engineering problems. Moreover, it allows for the recognition of the external variables and attitudes of non-computing engineers towards using data science in their work. We apply the framework in the study design, choosing our approach with Reddit in alignment with being able to understand individual and collective perceptions of data science, and interpretation.

3. Background

The foundation of this work is built upon the intersection of data science principles and web scraping methodologies applied to Reddit forums. We first begin with an overview of what data science is and how it may be applied in engineering (Section 3.1) to demonstrate why it is significant as a distinct field of study. Following that, to provide a foundation for our approach, we explain what web scraping is and some commonly used technologies to create the groundwork for the research in Section 3.2. Understanding this work necessitates a grasp of data science's significance and a foundational understanding of web scraping tools and techniques. Moreover, to raise awareness of how to practically analyze the data gathered, we elaborate on an overview of Natural Language Processing in Section 3.3.

3.1 Overview of Data Science

Data science, as defined by Provost and Fawcett, is “a set of fundamental principles that support and guide the principled extraction of information and knowledge from data” [1, p. 2]. Data science involves the use of statistical and computational methods to gather, analyze, and interpret large volumes of structured and unstructured data to inform decision-making, identify patterns, and make predictions. Data science involves several stages, including data collection, data preprocessing, data exploration, model building, model evaluation, and deployment. Various tools and techniques may be involved in these stages. In industries such as finance, healthcare, marketing, and technology, data science is used to improve business operations, improve customer experiences, and inform strategic decision-making [2], [3], [4], [5].

Due to the growing importance of data in modern engineering applications, data science principles are becoming more popular among engineers [19]. Engineers use data science techniques to analyze large volumes of data and extract insights that can inform design decisions, improve product performance, and improve processes. Data science is used to improve software design and user experience [20]. In mechanical engineering, for example, data science can be used to analyze sensor data from machines to detect anomalies, predict failures, and improve maintenance schedules [21]. Civil engineering uses data science to design more efficient transportation systems [22]. Chemical engineering as a field is currently shifting from being a discipline driven by “empirical and heuristic” principles to being more driven by artificial intelligence (AI) methods [23].

Machine learning and data science are interconnected disciplines that play crucial roles in various industries, including energy and aerospace. Machine learning is a subset of AI that enables computer systems to learn and improve from experience without being explicitly programmed [24]. To put it in better words, machine learning “is a branch of AI that uses algorithms to give robots the ability to learn from data and get better over time [25, p.1].” It involves the development of algorithms and statistical models that allow machines to recognize patterns in data and make predictions or decisions based on that information. For example, in the

context of the energy industry, engineers leverage machine learning to optimize power generation and distribution systems. By analyzing vast amounts of historical data from power plants, weather patterns, and consumption trends, machine learning algorithms can identify patterns and correlations that may not be apparent to human operators. These insights enable engineers to enhance power generation efficiency, reduce waste, and better match electricity supply with demand, ultimately leading to cost savings and reduced environmental impact [26].

Similarly, in the aerospace industry, machine learning is revolutionizing aircraft design and performance [27]. Engineers can use machine learning algorithms to analyze aerodynamic data, structural simulations, and historical performance data of aircraft to identify design improvements. These improvements may include more streamlined shapes, better materials selection, or better wing configurations, resulting in increased fuel efficiency, reduced emissions, and improved safety [27]. In Engineering Education, from an application point of view, natural language processing techniques have been used in broader applications such as facilitating student feedback [28], [29]; analyzing qualitative data to extract sentiments, emotions, and concerns [30]; theoretical framework selection [31]; analyzing mission statements [32] and so on.

Overall, these approaches and applications of data science in the context of engineering are present in the literature, however, getting a wide range of responses from the public forum outside the scholarly literature might offer deeper insight and could provide valuable understanding to the educational community. We take a systematic approach to gather more candid responses on perceptions of data science. Although engineering may encompass many subfields, we deliberately sought to keep the definition broader to allow for perceptions across sectors, industries, and institutions.

3.2 Overview of Web Scraping

In the literature, web scraping is defined as the “automatic retrieval of data from the Web for industry and academic research projects” [28, p.1], or the “procedure of automatic extraction of data from websites using software” [34], or “an interactive method for website and some other online sources to browse for and access data” [35]. Other definitions also extend these definitions by suggesting the collection of unstructured data from the web into structured ones in “a central database or spreadsheet” [36]. Web scraping is also referred to as web crawling, but some argue that web scraping is the extraction of data from a website, whereas web crawling is the identification of target Uniform Resource Locator (URL) links [34]. Broucke et al. extend on this and suggest that the crawling term refers to the ability of the program to navigate web pages on its own with the possibility of navigating with no precise end aim or purpose, continually exploring what a site or the internet has to offer [37].

There are many software resources available to configure web crawling techniques, which can make web scraping and data extraction much easier and faster. These tools may instantly detect a webpage's structure or provide a documentation framework that eliminates the need to manually generate web scraping code or other parsing functions that gather and convert data, as well as spreadsheet apps that store the extracted data [37]. Some popular tools include Scrapy [38], Beautiful Soup [39], Octoparse [40], ParseHub [41], and Application Programming Interface (API)-based tools [42]. Among these, APIs are helpful as through this, users can access their data directly without manual scraping. As defined by Amazon Web Services (AWS), in the context of APIs, the term Application refers to any program that performs a certain purpose of the interface that allows the contract of service between two apps. The term of the aforementioned contract specifies how the two will communicate with one another via requests and answers [42]. For example, Twitter API allows the retrieval of tweets and user data [43], and Google Maps API allows location and direction retrieval [44]. In the field of engineering education, researchers have used some of these tools to mine Industrial Engineering job postings (Beautiful Soup [45]); create a MOOCLink (combination of API-based tools and Scrapy [46]), and so on. Overall, the choice of tool depends on the specific needs and skills of the task or project, as well as the complexity of the website and the data that one wants to extract.

3.3 Overview of Natural Language Processing

Textual analysis uses computer systems to read and understand human-written text to generate insights [47]. Textual analysis uses natural language processing (NLP) to help classify, sort, and extract information from the text to find patterns, correlations, sentiments, and other actionable knowledge from unstructured textual data [48]. To facilitate NLP approaches, the Natural Language Toolkit (NLTK) [49], a Python library, is widely used.

There are multiple steps to build an NLP pipeline, each step with a brief overview is given below.

3.3.1 Data Preprocessing

This step is useful to turn raw data into a suitable form to carry out further analysis. This includes stopword and Noise removal and tokenization/Lemmatization. Stopwords are common words that do not add much meaning to a sentence, such as “the,” “and,” “of,” and “to”[50]. It is considered appropriate to remove such terms from the data to reduce the size of the dataset and enhance the predictive abilities of the algorithms employed [51]. Eliminating punctuation, special characters, and other irrelevant elements is known as noise removal. Tokenization refers to the process of splitting the text into individual words or tokens [52]. Lemmatization is the process of reducing words to their base or root form; taking the meaning of the word into account. For example, the words “walk,” “walked,” and “walking” would all be lemmatized to the lemma “walk.” Lemmatization helps to preserve the meaning of the words.

3.3.2 Data Analysis

After preprocessing, the raw data is transformed into a simple and standardized text, making it easy to further analyze. Two commonly used analytical techniques to extract information from data are N-gram and Latent Dirichlet Allocation (LDA) [53].

An N-gram [54] is a contiguous sequence of n items from a given sequence of text or speech. For example, for the sentence - "There was heavy rainfall," a 2-gram is a sequence of two words, such as "heavy rainfall". A 3-gram is a sequence of three words, such as "was heavy rainfall." N-grams can be used to predict the next word in a sentence or to identify patterns in text.

Among other textual analysis techniques, topic models have become increasingly useful in recent years as they are particularly helpful for tasks such as clustering documents, organizing large amounts of text data, and extracting information from unstructured text [55], [56]. Topic modeling can provide valuable insights and enhance our understanding of complex text data. For example, one study by Blei et al. proposed Latent Dirichlet Allocation, a generative probabilistic model for topic modeling [53]. LDA assumes that each document is a mixture of topics and that each topic is a distribution over words. The model uses Bayesian inference to estimate the topic distributions of each document and the word distributions of each topic. The authors demonstrated the effectiveness of LDA on several text corpora, including news articles and scientific papers [53].

4. Methods

To answer our RQ, we utilized the process described in Figure 2, which is similar to the approach described by Lunn et al. [57]. For this project, we divided our data handling and analysis approach into three parts, namely: 1) Data preprocessing, 2) Topic Modeling, and 3) Model Optimization and Tuning.

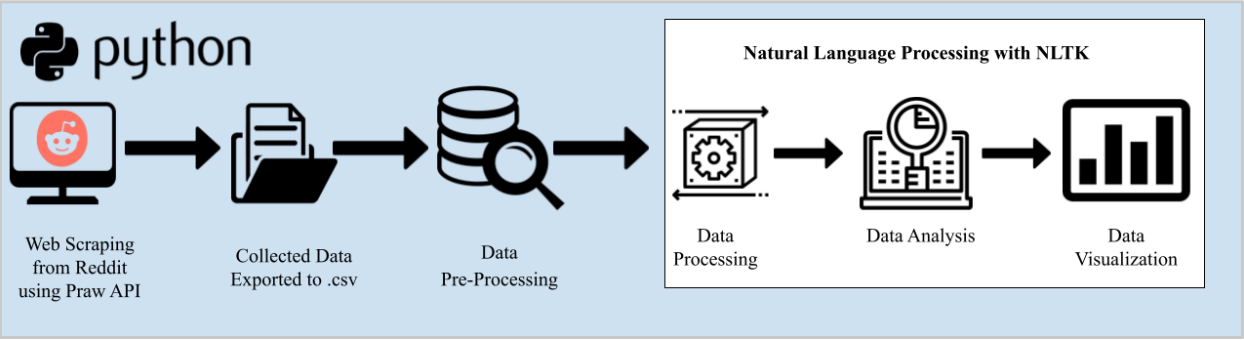

Figure 2. Overview of exploratory data collection and analysis process using PRAW and NLTK

In the section that follows, we first present an overview of the dataset itself and the data collection process. Then, we share more information about our approach to data handling and

analysis, including preprocessing, topic modeling, and model optimization. The code for this project can be provided upon request.

4.1 Data Collection

In this project, we elected to use an API-based tool for web scraping. Specifically, we used the Python Reddit API Wrapper or PRAW, which "is a Python package that allows for simple access to Reddit's API…it aims to be easy to use and internally follows all of Reddit's API rules" [58]. We picked Reddit as, in addition to being easily accessible, it was identified as a good source of data due to its high rate of daily users of 52 million [59], about 430 million active monthly users [60], and in contrast to other social media platforms, it also enables users to create subreddits, or smaller communities with pages devoted to particular topics [60], thus compartmentalizing the information for specific topics, which is valuable for data collection. A subreddit is defined as a section of Reddit that is a theme-based section of a forum. The platform allows the public to express themselves openly and provides a high level of anonymity [61], [62], thus making the information available there very valuable for research purposes. Using the PRAW API, we collected the data as described in the data collection section described below.

Within Reddit, we identified some subreddits pertaining to non-computing focused engineering. For example, "if you want to talk about gaming, you can go to r/gaming, but if you want to talk about a specific game like League of Legends, you go to r/leagueoflegends [63];" the r/subreddit is how the subreddit is identified under the platform. Here are some subreddits that we chose after verifying that they pertained to engineering: r/Engineering, r/EngineeringGradSchool, r/EngineeringResumes, r/civilengineering, r/Mechanicalengineering, r/ElectricalEngineering, r/Biomedicalengineering, r/Chemicalengineering, r/Industrialengineering, r/firePE, r/EnvironmentalEngineer, r/AerospaceEngineering, r/AskEngineers, r/LearnEngineering, r/aerospace, r/EngineeringStudents, r/AskanEngineer, and r/AskEngineers. We then specified the key term "data science" to capture all the instances where this phrase was used. After this search query was passed to the PRAW library's search function, it returned any content from Reddit that matched the search term "data science." Using the .csv module [64], the gathered data was written into a spreadsheet. We then processed the data as described in section 4.2 below. The final data collected included 20281 posts from 13 different subreddits. Characteristics of the dataset are presented in *Table 1*. As shown, for each subreddit, we incorporated both the overall count and unique instances of submission identifiers, submission titles, submission text, and comments.

4.2 Data Handling and Analysis with Natural Language Processing

4.2.1 Data Preprocessing

As part of the data preprocessing process, we did the following: 1) Expanded the contractions; 2) Remove URL as well as non-ASCII values (e.g., â€™), if any; 3) Remove punctuations; 4)

Tokenization - Converting a sentence into a list of words; 5) Remove stopwords; 6) Lemmatization /Stemming Transforming any form of a word to its root word. Some of the tasks are briefly discussed below:

First, we imported the inbuilt list from the NLTK. Stopwords. Then, we iteratively extended or removed stopwords as we went through the analysis, including the complete list provided in **Figure 3**. This list of custom stopwords was selected based on inconsistencies in the conversation that detracted from the overall understanding of the RQ and to minimize the corpus, which can also serve to enhance the training and testing time. Given that Reddit language is informal, such words represented typos, Reddit-specific jargon and requests, discipline-specific terms, those with no meaning, and greetings that did not contribute to the topic of focus.

Subreddit Name	**Submission Id**		**Submission Title**		**Submission Text**		**Comment**	
	Count	*Unique*	*Count*	*Unique*	*Count*	*Unique*	*Count*	*Unique*
AerospaceEngineering	1047	17	1047	17	1047	17	1047	179
AskEngineers	9180	92	9180	92	9180	92	9180	873
Chemicalengineering	1674	31	1674	31	1674	31	1674	336
Civilengineering	802	18	802	18	802	18	802	178
ElectricalEngineering	493	19	493	19	493	19	493	87
Engineering	380	7	380	7	380	7	380	94
EngineeringResumes	898	35	898	35	898	18	898	142
EngineeringStudents	2464	64	2464	64	2464	62	2460	557
EnvironmentalEngineer	18	1	18	1	18	1	18	6
Industrialengineering	543	24	543	24	543	24	543	121
Mechanicalengineering	1511	20	1511	20	1511	20	1511	314
aerospace	469	7	469	7	469	7	469	69
civilengineering	802	18	802	18	802	18	802	178
Total	**20281**	**353**	**20281**	**353**	**20281**	**334**	**20277**	**3134**

Table 1. Characteristics of the data collected from Reddit

```
stop_word.update(['like','would','really','also','x200b','could','since','thanks','thank',
                  'please','subredditmessagecomposetorengineeringstudents','pretty','much','yeah',
                   'doe','probably','rengineering', 'rengineeringstudents',
                  'later', 'might', 'year', 'hello', 'even','though', 'already',
                  'data','science','engineering', 'start','back','else','either', 'want','feel','sure',
                  'println','broken','wiki','read', 'rule','Adrian_Newey', 'beep','boop','whether',
                  'break', 'case', 'serial',
                  'moderator','page', 'well','turned'
                  'automoderator', 'accordingly',
                  'hippopotamusna'])
```

Figure 3. The extended list of stopwords added as we progressed through the analysis

To tokenize the corpus and lemmatize the tokens, we used NLTK packages, namely word_tokenize and WordNetLemmatizer [49].

To create an overview of the n-grams present in the data set, we utilized the n-gram package from the NLTK. The package also provides the frequency of the n-grams. After an iterative refinement process, we decided that implementing bigrams would be the most effective approach for our use of N-grams. To train the n-gram model, we have used Maximum Likelihood Estimation (MLE). The issues of MLE, and "zero probability N-grams" are well-documented [65]; however, we have not addressed this issue in this round and intend to optimize the model (e.g., employing smoothing techniques) further as part of our future work.

4.2.2 Topic Modeling

Latent Dirichlet Allocation is a widely used approach for topic modeling, and we used Python's Gensim as it provides a convenient and efficient way to implement it. The challenge with LDA is to find meaningful, clear, and good-quality topics that are heavily dependent on context and the pre-processed data. Therefore, to establish a baseline, we trained the LDA model with the following parameters, *num_topics=10, random_state=100, chunksize=100, passes=10, and per_word_topics=True*. To visualize the topics generated from the data, we used pyLDAvis, an interactive Python package.

4.2.3 Model Optimization and Tuning

To find the optimum number of topics, we built many LDA models with different values corresponding to the number of topics ranging from 2-11 and adjusted the value of alpha, α, and beta, β. We adopted this approach from Shashank Kapadia's article titled "Evaluate Topic Models: Latent Dirichlet Allocation (LDA): A step-by-step guide to building interpretable topic models [66]." Upon observing the results, we noted that the coherence score kept decreasing with the number of topics. Therefore, we picked K=3 as Shashank suggests that "it may make better sense to pick the model that gave the highest CV before flattening out or a major drop [66]." As for optimal α and β, we selected α=0.61, and β= 0.91 based on the highest CV, yielding approx. 16% improvement over the baseline score.

Topics	α	β	Coherence
3	0.61	0.91	0.62285136
3	0.61	symmetric	0.62285136
3	0.61	0.01	0.61553173
3	0.91	0.01	0.60897556
3	0.91	0.61	0.60895453

Table 2. The coherence score based on multiple combinations of α and β for the topic

4.2.4 Thematic Analysis

Using the keywords based on their Intertopic Distance Map with K=3 that were identified through the NLP-techniques, the first two authors referred back to the original dataset to gather information about the context for these terms. Then, they manually performed a thematic analysis. Others have previously used this approach in the context of engineering education. For example, Kardam et al. [67] have utilized thematic analysis to organize the collection of words assembled by LDA and then convert these topics/codes into themes to understand students' perception of the support provided by the Teaching Assistant. They independently assessed all the posts and established tentative categories. They then met to negotiate and refine.

5. Results

Top N-grams

The top 50 bi-grams extracted from the dataset are given in **Table 3**.

#	Bigram	Frequency	#	Bigram	Frequency
1	(machine, learning)	9716	26	(degree, industrial)	1139
2	(entry, level)	3366	27	(learning, actually)	1138
3	(make, sense)	2838	28	(academic, skill)	1132
4	(automation, engineer)	2417	29	(advance, advice)	1132

5	(high, tech)	2288	30	(towards, something)	1127
6	(mathematical, modelling)	2184	31	(degree, take)	1127
7	(programming, skill)	2152	32	(getting, another)	1126
8	(change, career)	1699	33	(skill, least)	1123
9	(applied, mathematics)	1389	34	(career, prospect)	1122
10	(bachelor, degree)	1354	35	(different, discipline)	1119
11	(construction, site)	1288	36	(previous, degree)	1116
12	(statistic, interested)	1247	37	(known, company)	1116
13	(need, know)	1229	38	(something, good)	1116
14	(deep, learning)	1222	39	(starting, career)	1114
15	(career, bachelor)	1221	40	(marketing, financial)	1106
16	(domain, knowledge)	1221	41	(energy, environmental)	1106
17	(spare, time)	1215	42	(signal, processing)	1104
18	(mechanical, engineer)	1197	43	(problem, solved)	1104
19	(linear, algebra)	1194	44	(skill, important)	1103

20	(research, paper)	1188	45	(another, bachelor)	1102
21	(renewable, energy)	1185	46	(need, change)	1101
22	(interesting, problem)	1162	47	(sound, exciting)	1101
23	(computer, programming)	1158	48	(problem, quite)	1101
24	(career, without)	1145	49	(company, getting)	1101
25	(engineer, professional)	1142	50	(job, focus)	1101

Table 3. Top 50 bi-grams

LDA Topic Modeling

Figure 4 gives topic visualization in the form of Intertopic Distance Map, where each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent that topic is. A good topic model will have fairly big, non-overlapping bubbles scattered throughout the chart instead of being clustered in one quadrant. A model with too many topics will typically have many overlaps, small-sized bubbles clustered in one region of the chart.

We wanted to tune model parameters and the number of topics to minimize circle overlap (Table 2). To understand how good or bad an LDA model is, there are multiple evaluation approaches, among them, one is Topic Interpretability. Topic coherence can be a useful metric for evaluating the effectiveness of a topic model and can guide the selection of hyperparameters and tuning of the model to produce more coherent and interpretable topics. The coherence score for the ten topics in this baseline model is 0.53. We used this score to tune the model by adjusting the hyperparameters alpha and beta. Also, after optimizing the model, we decided on to set the number of topics to K=3. The final model output is given in the form of an Intertopic Distance Map (*Figure 5*).

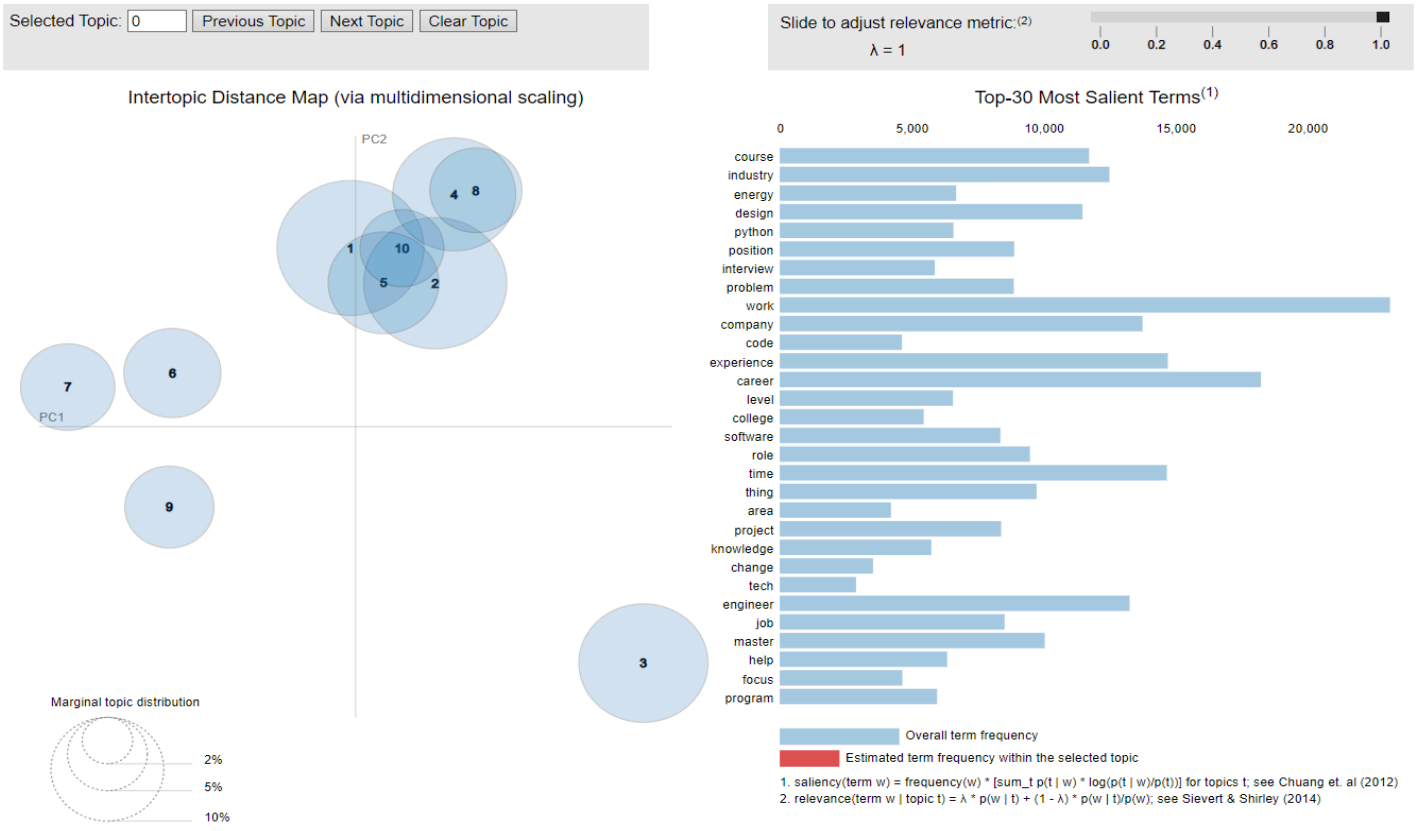

Figure 4. Intertopic Distance Map, with K=10

	0	1	2	3	4	5	6	7	8	9
Topic 1	interview	engineer	company	position	thing	experience	automation	work	internship	time
Topic 2	master	field	role	career	program	experience	course	level	research	class
Topic 3	career	energy	problem	knowledge	computer	work	focus	advice	machine	research
Topic 4	math	class	work	problem	course	system	school	time	analysis	field
Topic 5	work	skill	time	company	study	degree	thing	experience	industry	master
Topic 6	work	project	experience	matlab	time	company	engineer	industry	career	friend
Topic 7	industry	design	work	company	python	time	college	tech	software	experience
Topic 8	course	python	machine	code	help	language	certification	time	software	engineer
Topic 9	job	position	level	experience	pump	hour	change	role	school	construction
Topic 10	area	work	thing	engineer	job	path	offer	market	college	life

Table 4. Top 10 words in each of the topics

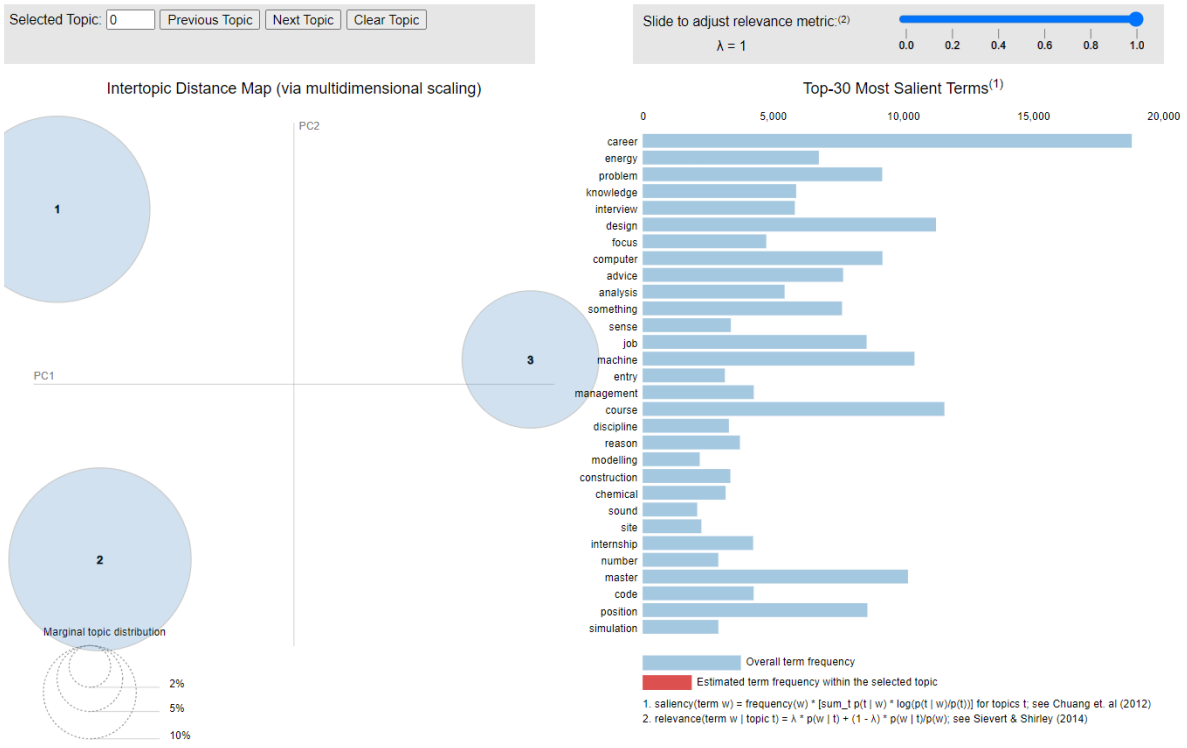

Figure 5A. Intertopic Distance Map with K=3, where each bubble on the left-hand side plot represents a topic.

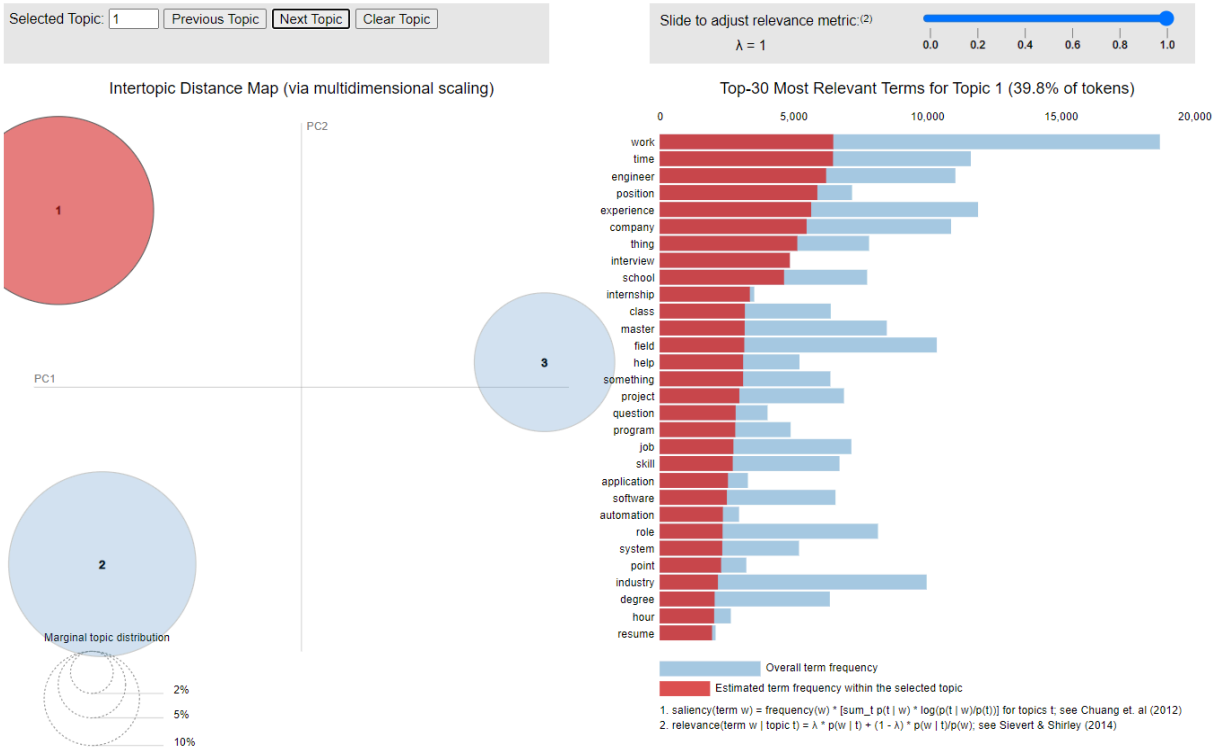

Figure 5B. Intertopic Distance Map for topic 1

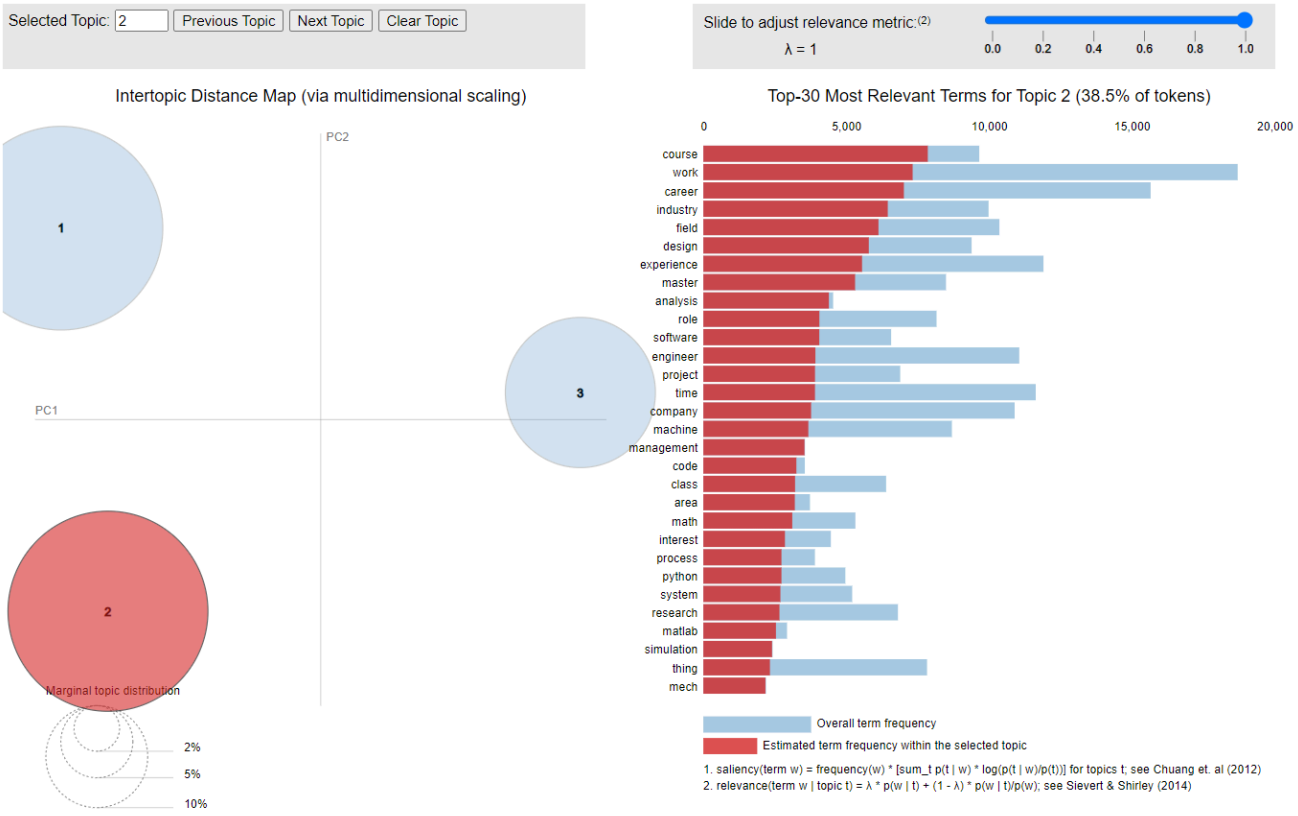

Figure 5C. Intertopic Distance Map for topic 2

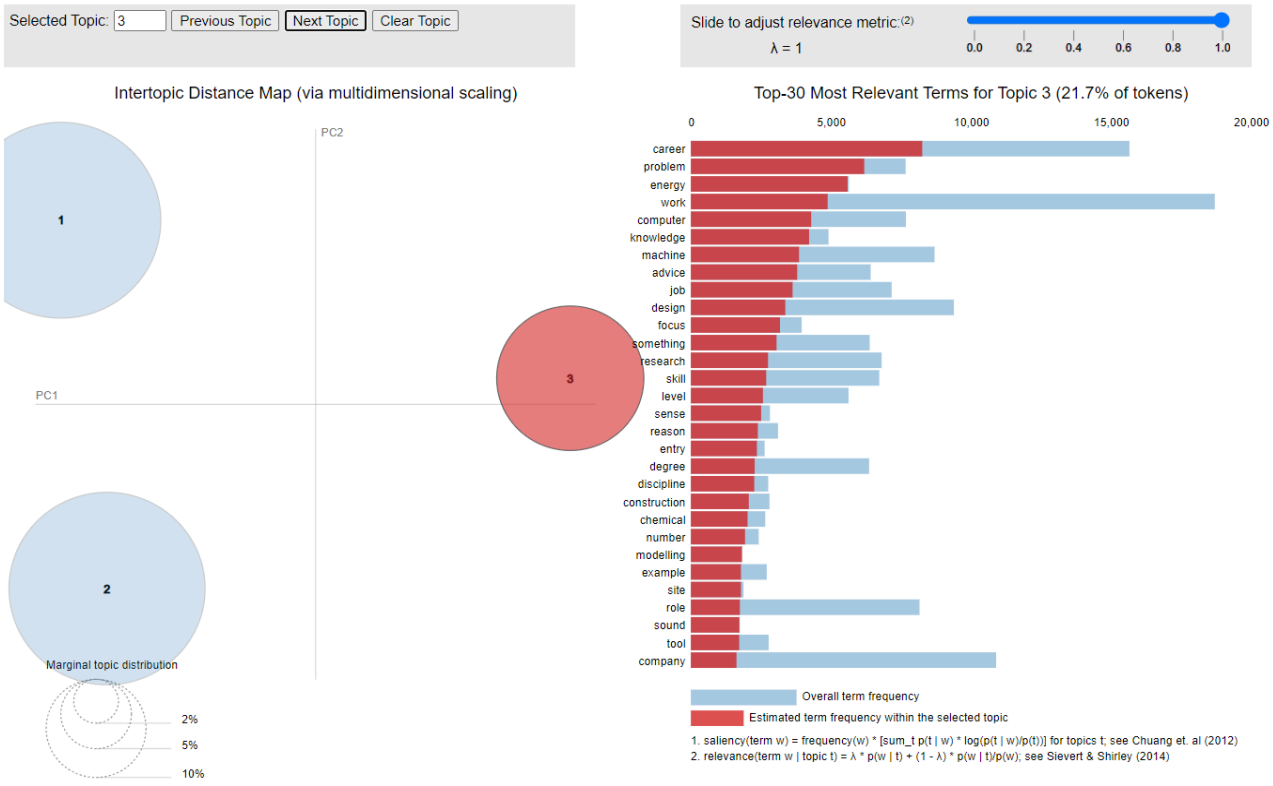

Figure 5D. Intertopic Distance Map for topic 3

Topics Extraction

In this study, we employed the Technology Acceptance Model as a theoretical framework to guide our inquiry into the use and acceptance of data science principles among non-computing engineers in approaching engineering problems. To infer the topics using the top keywords found in the LDA model, we looked at the Reddit posts in the data set to better understand the context. This was conducted manually, and then we applied thematic analysis [68] to extract a broader categorization around the RQ. Our analysis identified three main topics, which we labeled as "Job search, undergrads, and professional development," "Continuous Education and Career Upskilling," and "Multiple Applications," as seen in **Figure 6**.

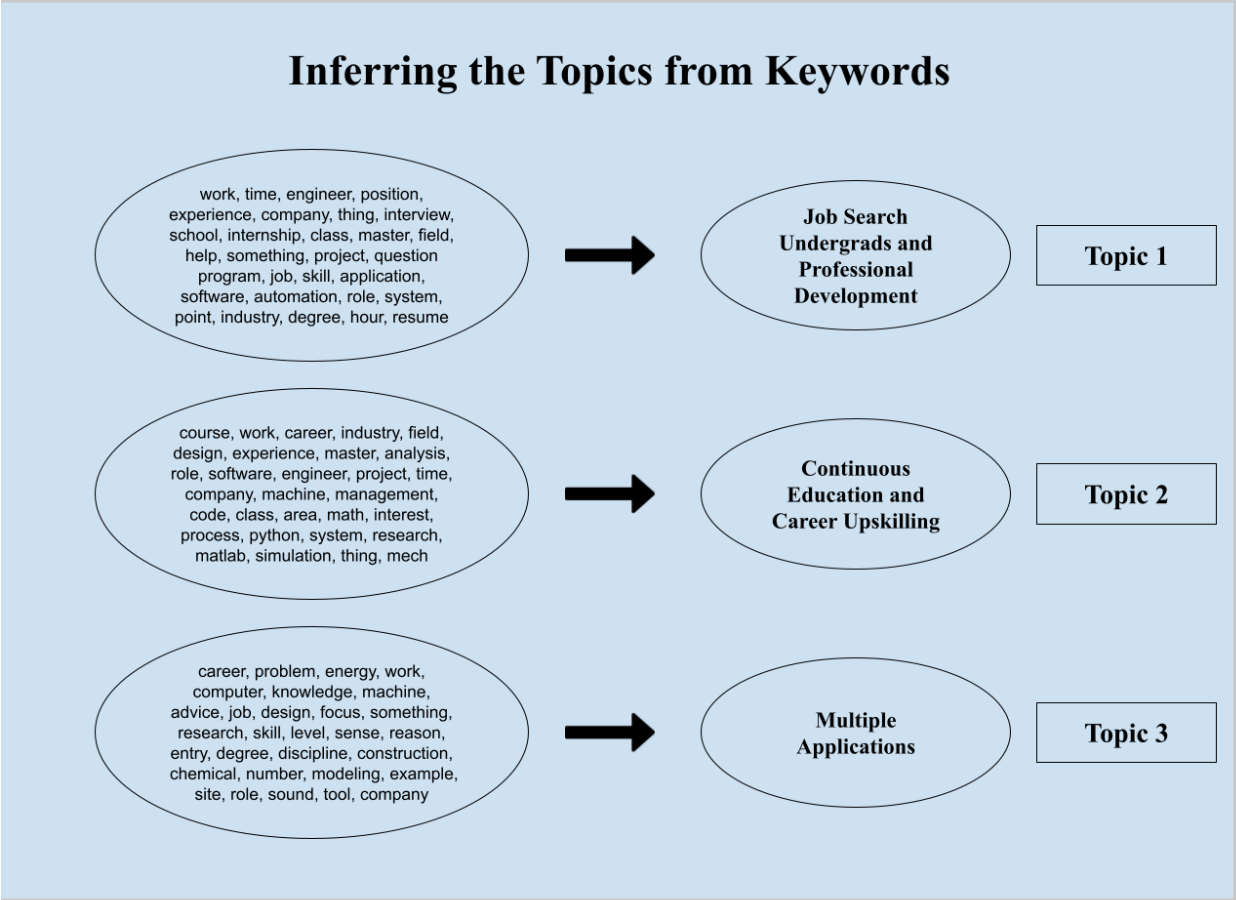

Figure 6. Keywords mapping based on topic modeling performed on the data

The first topic is more **career-oriented** and contains these specific keywords such as work, time, engineer, position, experience, company, interview, school, internship, project, program, job, skill, application, software, automation. Overall, the keywords seem to capture potential attitudes relating to the role and fit of data science in engineering work, required abilities, education and training, career implications, and general considerations around its perceived usefulness and applicability.

Words like “work,” “time,” “experience,” “position,” “company,” “project,” “job” might suggest examining perspectives on how data science fits into engineering roles and day-to-day work. Terms such as “engineer,” “field,” “industry,” “role,” “system” might imply studying views on data science's place within the broader engineering profession. Concepts like “skill,” “software,” “automation,” and “system” might indicate attitudes related to technical abilities and computing skills needed for data science may be relevant. Keywords around education and training like “master,” “class,” “program,” and “degree” might suggest studying opinions on preparation and qualifications needed for data science. Words such as “resume,” “application,” “interview,” and

“hire” might point to considerations around career development and advancement in data science roles. Phrases like “help,” “something,” “question,” and “point” might imply perspectives on the value and application of data science may be examined.

As an example, under the r/ElectricalEngineering subreddit forum, using the keyword “position,” a user submitted the following:

> *Hello all! I need some advice between these two majors. I'm currently enrolled and ready to start ASU online Electrical Engineering program. Some future jobs I think would suit me and I have interest in are as follows. -Computer Vision/ML -Digital Signal Processing -Embedded Programming -Software Engineering of all flavors -Data science or bioinformatics (Most of these positions I see myself working more on the software side although I do have a soft spot for some hardware applications). Which do you think is a better major to cover these prospects? I am thinking of going the route of BSEE to MSCS, although I'd like to still have good options if I don't pursue a graduate degree. COMPUTER ENGINEERING IS NOT AN OPTION. NEITHER IS A EE MAJOR WITH CS MINOR unfortunately -__-*

Similarly, under the r/Chemicalengineering subreddit forum, using the keyword “internship,” a user submitted the following:

> *Hi, I'm a junior in ChemE and I've been applying to internships over winter break. I noticed that a lot of battery engineering companies are quite small or at a startup level. The main exception seems to be the automotive industry. I'm wondering what the future of battery engineering in the US will be like. Is it actually projected to grow as fossil fuels begin to phase out? How much will this industry grow and how secure will jobs be? In addition, I've heard that a lot of large battery companies like Samsung will continue to have large plants in Asia, and battery engineering is significantly more popular in countries like China and Korea rather than in the US. Is there any truth to this, and will this impede the growth of the industry in the US?*

To which another user replied:

> *Industries tend to follow an S-curve (Sigmoidal curve), where in their birth they expand very slowly, undergo a massive expansion period, and then plateau towards the end…It's not important to take courses in this field as a bachelor's IMO because the PhDs will always beat you to the jobs, unless you have more internships/work experience than them. A lot of what you do will probably be learned on the job because few schools teach electrochem in the perspective of new, evolving cell chemistries…If you're interested in data science, test engineering or more specifically, cell testing and evaluation is definitely the path to go. There's a lot of number crunching to be done due to the nature of cell testing with large timeframes (like 10 days - 1 month of data, 20 parameters evaluated every*

second). Tesla actually has an entire team dedicated to cell & pack data analysis. I'm willing to guess it's the same for Apple. You could also explore Battery management systems, which controls the operating conditions on the pack level.

The second topic is more focused on **continuous education and career upskilling** and contains these specific keywords: course, work, career, industry, field, design, experience, master, analysis, role, software, engineer, project, time, company, machine, management, code, class, area, math, interest, process, python, system, research, matlab, simulation, thing, mech. The keywords overall seem to capture themes around perceptions of data science's place in the engineering profession, its relationship to technical skills, and considerations around training and career development in data science. The attitudes and viewpoints around these areas in non-computing engineering fields could be relevant to explore for the research. Several keywords relate to engineering work and careers more broadly, like "work," "career," "industry," "field," "design," "experience," "project," "company," and "management." This suggests the research may explore how data science fits into the overall engineering profession. There are keywords about specific engineering disciplines like "mech," "design," "simulation." This might suggest studying attitudes across engineering subfields. Some keywords reference engineering skills like "analysis," "math," "matlab," and "python." The research may look at perceptions of needed technical abilities for data science. Terms like "software," "code," "machine," "system" might indicate data science is seen in relation to programming and computing skills. The attitudes on this connection could be examined. Concepts like "role," "interest," "process" suggest investigating how data science is viewed regarding its purpose, appeal, and procedures in engineering work. Keywords about training like "master," "class," "course" imply attitudes related to education and preparation for data science may be studied.

As an example, under the r/Civilengineering subreddit forum, using the keyword "experience," a user submitted the following:

Hello I am curious if anyone in this sub can share the names of companies or niches within civil engineering where programming, statistical skills, and working with data comes into play. I have done my research and it appears that coastal engineering seems to be a niche where modeling is the closest to this. It also seems like a tough discipline to get your foot into the door without a PhD. Background: I am a civil engineer who is about to graduate with my MS. (I worked in the industry as a transportation engineer for 2 years between undergrad and grad school in order to gain experience and figure out my interests and innate skills). During my graduate degree, I focused on honing data science skills. I took courses such as machine learning, geospatial analysis, data science for environmental site characterization, hydrological modeling with python, etc. My research project was interdisciplinary and dipped into health care. I am open to a job far from engineering, but I love civil and its applications. I also am thinking towards the future and I see myself achieving my P.E., which wouldn't happen if I went to a healthcare tech firm. Thanks

Furthermore, under the r/AskEngineers subreddit forum, using the keyword “python,” a user submitted the following:

> *Hi! I'm a young mechanical/energy engineer, graduated with a BS in MechEng in 2015, have been laid off twice (second time was a few days ago), from the energy modeling/building energy engineering industry, and I'm looking for a change. I've always been decent at math, and while I had to work at it in college, I maintained A-B's at WPI (ok engineering school). I've been interested in data science and algorithms for a while, and took a couple Python courses on Coursera before starting my latest job. While I enjoyed it, I found it somewhat challenging.*

The third topic centers on **multiple applications**, including terms such as career, problem, energy, work, computer, knowledge, machine, advice, job, design, focus, something, research, skill, level, sense, reason, entry, degree, discipline, construction, chemical, number, modeling, example, site, role, sound, tool, company. The keywords seem to indicate studying attitudes relating to career impacts, usefulness for engineering work, required skills and training, and general considerations around the role and value of data science across different non-computing engineering disciplines. Words like “career,” “job,” “company” might suggest examining attitudes relating to how data science impacts engineering career paths and job prospects. Terms such as “problem,” “design,” “research,” “modeling” might imply perspectives on the value and applicability of data science for engineering work like problem-solving, research, and design. Concepts like “energy,” “chemical,” “construction” might indicate studying views across non-computing/specialized subfields of engineering. Keywords around skills like “computer,” “knowledge,” “skill,” “tool” might point to attitudes regarding technical abilities needed for data science. Words such as “entry,” “degree,” “level” might suggest considerations around qualifications and training required to get into data science. Phrases like “advice,” “focus,” “reason,” “example” might imply examining perspectives on the role and fit of data science within engineering work. Terms including “sound,” “sense,” “something” may capture general sentiments, whether positive or negative, towards data science.

As an example, under the r/Mechanicalengineering subreddit forum, using the keyword “chemical,” a user submitted the following:

> *…Valves, sensors, pumps, everything's getting more and more connected as iot and automated. Refineries can respond immediately to shifts in concentration of hazardous chemicals in ppm and adjust the entire process on the fly. Machine learning in general helps us build models for systems from experimental data rather than having to correct theoretical models and hope the accuracy stays good.*

This was an answer to the following initial post:

> *I am trying to mesh my mechanical engineering degree with my data science/programming background to improve my chances of getting employed in companies that need someone who can do both. I am trying to do that by doing a project that is relevant to what the industry needs today but all I know is that Neural Networks are a hot topic right now in CFD. I don't know what those applications are or the problem statements they are trying to tackle. Can someone guide me to what the applications are and if possible also tell me where I can get the data to train and test my model?*

Overall, these quotes depict applications of data science principles in engineering where the users have some positive behaviors towards data science technologies.

6. Discussion

This study explored non-computing engineers' attitudes and considerations regarding the use of data science principles and technologies in engineering work through the lens of the Technology Acceptance Model. Our analysis of discussions in Reddit forums frequented by engineers revealed several key themes related to the perceived usefulness and ease of use of data science, which are core factors influencing technology acceptance in TAM. With that said, we noticed that the use of data science has broad implications in engineering, ranging from the optimization of engineering processes to the development of new products and technologies. The optimization of processes is one of the main uses of data science principles in engineering [69], [70].

The posts on Reddit often highlighted posters' attitudes towards using data science, as mentioned by the TAM. Often, Reddit posters described how data science principles could be applied to help solve a diversity of problems (Topic 3). This aligns with TAM's concept of perceived usefulness, wherein individuals are more likely to adopt a technology if they perceive it as beneficial and advantageous. However, they also saw learning data science as vital and part of their continuous education and career upskilling (Topic 2), suggesting that they viewed learning data science as vital and relatively accessible. This resonates with the TAM's perceived ease of use factor, which posits that individuals are more willing to adopt a technology if they perceive it as effortless or straightforward to use; in general, the Reddit users described it in terms of courses and work, but also research experiences. Furthermore, they mentioned how it could be important for jobs they were interested in, or their professional development (Topic 1).

The example described above, about a junior chemical engineering student looking for an internship related to "battery engineering," is interesting, as more electric vehicles (EVs) and eco-friendly devices reliant on battery technologies are entering the market. Engineers can find process inefficiencies and make data-driven decisions to fix them by analyzing massive datasets. For instance, in the design of lithium-ion batteries, predictive modeling is one of the ways to

efficiently accelerate the implementation of these technologies, and its improvement is powered by data-driven modeling, namely by "machine-learning model development" [69]. Another example is the use of data science to optimize traffic flow in urban areas. Razali et al. give a thorough grasp of how machine learning and deep learning approaches may be applied to enhance traffic flow prediction, boosting intelligent transportation systems in smart cities [22]. In water resources engineering, data science has been used to develop predictive models for flood risk management. Mosavi et al. did a literature review on using machine learning modeling to predict floods and found that it is still in its infancy and undergoing development [71]. Consequently, these examples are good illustrations of how machine learning is being used in engineering and the common trends in the industry.

The second top bi-gram was "entry level." This phrase highlights the challenges and experiences faced by recent graduates and young professionals, as well as for those entering the job market for the first time. Engineering is a field that requires constant learning and adaptation to new technologies. Novice job seekers in this area may face numerous challenges and discouraging experiences [72], [73]. For example, Mabey [74, p.1] stated that "there is a mismatch between graduate talent and the requirements of industry…graduate turnover is as high as 50 percent from first employers after 5 years. The problem is particularly acute in the engineering sector where there is difficulty first attracting and then retaining good quality graduates." In addition, some studies suggest that there is a gap between STEM education and the skills needed in the industry [75], [76]. In order to bridge that gap, students take their learning into their own hands and use places like Reddit to connect with other people to get informed and better their skill set.

Within the "entry-level" bi-gram, we also found a significant portion of a group of individuals who mentioned looking to transition from engineering to data science-related jobs. We did not see this as a trend that was often discussed in the literature. This represents an area for possible future exploration and a need for enhanced understanding. There are some articles that provide some steps or strategies to help with this transition. For example, M. Khorasani on Medium [77] stated the following:

> *You will indeed be able to transition from engineering to data science, but it will come through with impeccable perseverance, a small yet tangible setback in your career (as you jump branches), and a strict regiment of discipline. As you progress upwards on the corporate data science ladder, you should move from one position to another…You will become a hybrid of a data scientist and an engineer with the best of both worlds and you will take pride in knowing that you belong to a rare breed of professionals with a multidisciplinary skill set that should be of great value to most employers.*

Other frequent bi-grams, such as "renewable energy," "marketing financial," "energy environmental," "mathematical modeling," and "programming skill" suggest a growing demand for professionals with expertise in these areas.

7. Implications

Our study has several important implications for engineering education and practice. The findings highlight the need for greater integration of data science principles across engineering curricula, not just in computer science programs. Data science skills are becoming increasingly relevant across engineering disciplines, so all students should have opportunities to develop basic competencies. This may require revising program requirements, developing new cross-disciplinary courses, and/or providing faculty training on effective pedagogies for teaching data science topics, since they may be unfamiliar with them on their own.

For educators, the findings identify several critical skill gaps and training needs that should be addressed in engineering courses. More data science, programming, and mathematical modeling courses might assist students in getting the technical skills required for data-driven engineering positions. However, technical skills alone are insufficient. To meet industry requirements, courses should also focus on non-technical, also referred to as professional, skills such as communication, teamwork, problem solving, and other competencies specified by ABET [78]. Capstone projects, internships, and industry partnerships may all provide excellent real-world experiences. However, universities that apply "industry-engaged" curricula [79] could give their students an edge. Thus, educators should regularly revise courses based on input from alumni and industry partners. While higher education is not the same as a trade school, it is critical to consider the variables seen as pertinent for students' employability and to incorporate ways to cultivate such competencies and skills.

For administrators and policymakers, the findings emphasize the need to boost STEM funding and initiatives, particularly at the K-12 level as raised by some educators [80], [81]. Increased access to high-quality STEM education early on can spark more student interest and prepare a robust engineering talent pipeline [82]. Administrators can also promote partnerships between universities, government labs, and private companies to facilitate knowledge sharing and the development of customized training programs. Tax incentives can motivate companies to invest in continuing education and upskilling of employees [83].

For engineers looking to transition into data science roles, the key is a learn-by-doing approach. This approach is based on the assumption "that a person either prefers to learn by seeing the subject matter, hear about the subject matter, or do a task related to the subject matter" [84], which is also in alignment with the TAM. Taking online courses in programming, data analytics, and machine learning is a good start to building relevant skills. However, hands-on projects and internships are essential to put skills into practice. Engineers should take advantage of open datasets and platforms like Kaggle [85] to work on realistic data challenges and gain practical experience. Joining professional associations, networking events, hackathons, and participating in industry conferences allows connecting with mentors and recruiters. Patience and persistence are vital, as the transition can take time. Leveraging an engineering background and developing

versatile skill sets (as mentioned in the theme of continuous education and career upskilling, topic 2) can open up many opportunities.

The results also demonstrate that many engineering students may have a strong interest in learning data science to enhance their career prospects. This suggests that universities could provide additional electives, certificates, or other flexible learning pathways in data science integrated with existing engineering curricula. Students are even turning to external online platforms like Coursera [86] and edX [87] to supplement gaps in their formal education, indicating current unmet demands.

For employers, our findings reinforce the need to provide ongoing professional development in data science, as engineering graduates often lack full proficiency. Investing in upskilling programs and on-the-job training in data science can help engineers apply these tools effectively in their work. However, clear incentives are needed to encourage adoption, and it is important for managers to recognize the time and overhead such efforts may necessitate.

Ultimately, a more concerted, coordinated effort is required to integrate data science into engineering. Collaboration between educators, employers, government agencies, and professional associations can help set standards around data science learning objectives, develop curricular frameworks, and provide resources to support implementation. More research is also needed to identify effective pedagogies, evaluate integration models, and assess long-term outcomes, and engineering as we know it becomes more interdisciplinary. Working together, and coordinating between academia and industry, we hope to build a shared vision for the future of data-driven engineering education.

8. Limitations and Future Directions

There are several limitations that we want to acknowledge. First, our data was collected using subreddits that we identified as pertaining to non-computing engineering fields. Although other subreddits were beyond the scope of the present study, we should point out that this is not an exhaustive list of what is contained on Reddit, and different options could result in variable outcomes. In the future, the addition of more subreddits could provide a broader look at the current state of engineers' understanding and applications, and provide new insight. Moreover, while anyone can post to Reddit, non-computing engineers using data science may also exist in other parts of the world and share their opinions and experiences through other sites or in other languages. Going forward, it may be worth expanding our analysis to include additional data sources, and potentially to consider posts in other written languages.

Additionally, our analysis focused solely on discussions from Reddit, which may not accurately represent actual curricula, syllabi, and student information within specific engineering disciplines. Analyzing primary source materials like course catalogs, syllabi, and student records would likely provide more direct and reliable insights into how data science concepts are being

taught and applied in various engineering programs. While the Reddit data offers a window into community discourse, a more comprehensive examination of institutional resources is warranted. It is important to note that our focus was not limited to just students, but also included practicing engineers and those looking to apply data science principles, either during their studies or in industry settings. Therefore, while Reddit provided an appropriate data source for this broader perspective, future work could narrow the scope to explore student experiences and education specifically through analysis of syllabi and related materials.

9. Conclusion

Based on the analysis of Reddit posts related to data science applications in engineering, several key conclusions can be drawn. First, data science is perceived as highly useful by non-computing engineers for enhancing engineering work and career prospects. The analysis revealed prevalent positive sentiment towards data science and its ability to optimize processes, drive innovation, and provide valuable insights across engineering domains. This aligns with the “perceived usefulness” dimension of the Technology Acceptance Model.

Second, while viewed as useful, data science is not yet seen as easy to implement in engineering contexts. The posts indicated concerns about required technical skills, the need for training, and challenges in adapting data science tools to engineering problems. This highlights issues around the “perceived ease of use” dimension of TAM that may hinder adoption. Further education and practical experience are necessary to increase engineers' comfort and proficiency with data science.

Third, applications of data science span multiple engineering subfields, with opportunities to enhance design, research, simulation, and more. However, the realization may depend on the development of tailored, domain-specific implementations rather than one-size-fits-all solutions. Customizing data science to the unique needs of each engineering discipline will be key.

Fourth, data science has significant career impacts for engineers. The analysis revealed great interest in developing data science skills to boost employability and transition to data-driven roles. This talent pipeline should be fostered through updated curricula, training programs, and collaboration between academia and industry.

Finally, while highlighting perceived benefits, we should also think about the potential negative consequences of increased data science integration in engineering. Ethical implications, overreliance on algorithms, and improper usage are risks requiring diligent oversight. A measured, responsible approach to adoption is prudent. Our study also demonstrates the utility of web scraping as a data collection method and of LDA in analyzing large text datasets and extracting meaningful information. As technologies evolve and data science becomes increasingly pervasive, it is crucial that individuals across all disciplines develop a fundamental

level of data literacy. Proficiency in knowledge and applications of data science will provide future engineers with a significant advantage; therefore, it is imperative that we explore effective methods to better equip engineers with the essential foundations they need to excel in their post-collegiate endeavors.

References

[1] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013, doi: 10.1089/big.2013.1508.

[2] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” *J. Big Data*, vol. 6, no. 1, p. 54, Jun. 2019, doi: 10.1186/s40537-019-0217-0.

[3] J. Huttunen, J. Jauhiainen, L. Lehti, A. Nylund, M. Martikainen, and O. M. Lehner, “BIG DATA, CLOUD COMPUTING AND DATA SCIENCE APPLICATIONS IN FINANCE AND ACCOUNTING,” 2019.

[4] P. Chintagunta, D. M. Hanssens, and J. R. Hauser, “Editorial—Marketing Science and Big Data,” *Mark. Sci.*, vol. 35, no. 3, pp. 341–342, May 2016, doi: 10.1287/mksc.2016.0996.

[5] G. Langlois and G. Elmer, “THE RESEARCH POLITICS OF SOCIAL MEDIA PLATFORMS,” *Res. Polit.*, 2013.

[6] L. Cao, “Data Science: A Comprehensive Overview,” *ACM Comput. Surv.*, vol. 50, no. 3, Art. no. 3, May 2018, doi: 10.1145/3076253.

[7] D. A. C. Beck, J. M. Carothers, V. R. Subramanian, and J. Pfaendtner, “Data science: Accelerating innovation and discovery in chemical engineering,” *AIChE J.*, vol. 62, no. 5, pp. 1402–1416, May 2016, doi: 10.1002/aic.15192.

[8] O. H. and K. Mike, “Ten Challenges of Data Science Education.” Accessed: Feb. 06, 2024. [Online]. Available: https://cacm.acm.org/blogs/blog-cacm/246219-ten-challenges-of-data-science-education/fulltext

[9] N. Leger and B. Berhane, “Work in Progress: A Literature Review On Computational & Numerical Methods in Engineering Education,” presented at the 2022 ASEE Annual Conference & Exposition, Aug. 2022. Accessed: Sep. 05, 2022. [Online]. Available: https://strategy.asee.org/work-in-progress-a-literature-review-on-computational-numerical-methods-in-engineering-education

[10] A. Karpatne *et al.*, “Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data,” *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2318–2331, Oct. 2017, doi: 10.1109/TKDE.2017.2720168.

[11] “J. Manyika, ‘Big data: The next frontier for innovation, competition, and productivity,’ McKinsey Global Institute (MGI), 11-May-2011. [Online]. Available: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Where%20machines%20could%20replace%20humans%20and%20where%20they%20cant/Where-machines-could-replace-humans-and-where-they-cant-yet.pdf. [Accessed: 11-Apr-2023].”

[12] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Q.*, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

[13] V. Venkatesh and F. D. Davis, “A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies,” *Manag. Sci.*, vol. 46, no. 2, pp. 186–204, Feb. 2000, doi: 10.1287/mnsc.46.2.186.11926.

[14] A. Granić and N. Marangunić, “Technology acceptance model in educational context: A systematic literature review,” *Br. J. Educ. Technol.*, vol. 50, no. 5, pp. 2572–2593, 2019, doi: 10.1111/bjet.12864.

[15] N. Marangunić and A. Granić, “Technology acceptance model: a literature review from

1986 to 2013," *Univers. Access Inf. Soc.*, vol. 14, no. 1, pp. 81–95, Mar. 2015, doi: 10.1007/s10209-014-0348-1.
[16] Y.-C. Ku, R. Chen, and H. Zhang, "Why do users continue using social networking sites? An exploratory study of members in the United States and Taiwan," *Inf. Manage.*, vol. 50, no. 7, Art. no. 7, Nov. 2013, doi: 10.1016/j.im.2013.07.011.
[17] A. Tarhini, R. Masa'deh, K. A. Al-Busaidi, A. B. Mohammed, and M. Maqableh, "Factors influencing students' adoption of e-learning: a structural equation modeling approach," *J. Int. Educ. Bus.*, vol. 10, no. 2, pp. 164–182, Jan. 2017, doi: 10.1108/JIEB-09-2016-0032.
[18] A. A. AlQudah, M. Al-Emran, and K. Shaalan, "Technology Acceptance in Healthcare: A Systematic Review," *Appl. Sci.*, vol. 11, no. 22, Art. no. 22, Jan. 2021, doi: 10.3390/app112210537.
[19] *Understanding the Educational and Career Pathways of Engineers*. Washington, D.C.: National Academies Press, 2018. doi: 10.17226/25284.
[20] M. Kim, T. Zimmermann, R. DeLine, and A. Begel, "Data Scientists in Software Teams: State of the Art and Challenges," *IEEE Trans. Softw. Eng.*, vol. 44, no. 11, Art. no. 11, Nov. 2018, doi: 10.1109/TSE.2017.2754374.
[21] J. Zou, Q. Chang, Y. Lei, and J. Arinez, "Production System Performance Identification Using Sensor Data," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 48, no. 2, pp. 255–264, Feb. 2018, doi: 10.1109/TSMC.2016.2597062.
[22] N. A. M. Razali, N. Shamsaimon, K. K. Ishak, S. Ramli, M. F. M. Amran, and S. Sukardi, "Gap, techniques and evaluation: traffic flow prediction using machine learning and deep learning," *J. Big Data*, vol. 8, no. 1, p. 152, Dec. 2021, doi: 10.1186/s40537-021-00542-7.
[23] V. Venkatasubramanian, "The promise of artificial intelligence in chemical engineering: Is it here, finally?," *AIChE J.*, vol. 65, no. 2, pp. 466–478, Feb. 2019, doi: 10.1002/aic.16489.
[24] "Machine learning, explained | MIT Sloan." Accessed: Jul. 25, 2023. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
[25] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," *Cogn. Robot.*, vol. 3, pp. 54–70, Jan. 2023, doi: 10.1016/j.cogr.2023.04.001.
[26] D. Rangel-Martinez, K. D. P. Nigam, and L. A. Ricardez-Sandoval, "Machine learning on sustainable energy: A review and outlook on renewable energy systems, catalysis, smart grid and energy storage," *Chem. Eng. Res. Des.*, vol. 174, pp. 414–441, Oct. 2021, doi: 10.1016/j.cherd.2021.08.013.
[27] S. L. Brunton *et al.*, "Data-Driven Aerospace Engineering: Reframing the Industry with Machine Learning," *AIAA J.*, pp. 1–26, Jul. 2021, doi: 10.2514/1.J060131.
[28] A. Katz, M. Norris, A. M. Alsharif, M. D. Klopfer, D. B. Knight, and J. R. Grohs, "Using Natural Language Processing to Facilitate Student Feedback Analysis," presented at the 2021 ASEE Virtual Annual Conference Content Access, Jul. 2021. Accessed: Feb. 06, 2024. [Online]. Available: https://peer.asee.org/using-natural-language-processing-to-facilitate-student-feedback-analysis
[29] I. Anakok, J. Woods, M. Huerta, J. Schoepf, H. Murzi, and A. Katz, "Students' Feedback About Their Experiences in EPICS Using Natural Language Processing," in *2022 IEEE*

Frontiers in Education Conference (FIE), Uppsala, Sweden: IEEE, Oct. 2022, pp. 1–9. doi: 10.1109/FIE56618.2022.9962557.

[30] A. Satyanarayana, K. Goodlad, J. Sears, P. Kreniske, M. F. Diaz, and S. Cheng, "Using Natural Language Processing Tools on Individual Stories from First-year Students to Summarize Emotions, Sentiments, and Concerns of Transition from High School to College," presented at the 2019 ASEE Annual Conference & Exposition, Jun. 2019. Accessed: Feb. 07, 2024. [Online]. Available: https://peer.asee.org/using-natural-language-processing-tools-on-individual-stories-from-first-year-students-to-summarize-emotions-sentiments-and-concerns-of-transition-from-high-school-to-college

[31] C. G. P. Berdanier, C. M. McComb, and W. Zhu, "Natural Language Processing for Theoretical Framework Selection in Engineering Education Research," in *2020 IEEE Frontiers in Education Conference (FIE)*, Uppsala, Sweden: IEEE, Oct. 2020, pp. 1–7. doi: 10.1109/FIE44824.2020.9274115.

[32] S. Bhaduri and T. Roy, "Demonstrating Use of Natural Language Processing to Compare College of Engineering Mission Statements," presented at the 2017 ASEE Annual Conference & Exposition, Jun. 2017. Accessed: Feb. 07, 2024. [Online]. Available: https://peer.asee.org/demonstrating-use-of-natural-language-processing-to-compare-college-of-engineering-mission-statements

[33] Murray State University, V. Krotov, L. Johnson, Murray State University, L. Silva, and University of Houston, "Legality and Ethics of Web Scraping," *Commun. Assoc. Inf. Syst.*, vol. 47, pp. 539–563, 2020, doi: 10.17705/1CAIS.04724.

[34] M. Khder, "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application," *Int. J. Adv. Soft Comput. Its Appl.*, vol. 13, no. 3, Art. no. 3, Dec. 2021, doi: 10.15849/IJASCA.211128.11.

[35] B. S. Manjushree and G. S. Sharvani, "Survey on Web scraping technology," *Wutan Huatan Jisuan Jishu*, vol. 16, no. 6, Art. no. 6, 2020.

[36] S. de S. Sirisuriya, "A Comparative Study on Web Scraping," 2015.

[37] S. vanden Broucke and B. Baesens, "From Web Scraping to Web Crawling," in *Practical Web Scraping for Data Science: Best Practices and Examples with Python*, S. vanden Broucke and B. Baesens, Eds., Berkeley, CA: Apress, 2018, pp. 155–172. doi: 10.1007/978-1-4842-3582-9_6.

[38] "Scrapy | A Fast and Powerful Scraping and Web Crawling Framework." Accessed: Feb. 21, 2023. [Online]. Available: https://scrapy.org/

[39] "Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation." Accessed: Feb. 21, 2023. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[40] "Web Scraping Tool & Free Web Crawlers | Octoparse." Accessed: Feb. 21, 2023. [Online]. Available: https://www.octoparse.com/

[41] "ParseHub | Free web scraping - The most powerful web scraper." Accessed: Feb. 21, 2023. [Online]. Available: https://www.parsehub.com/

[42] "What is an API? - Application Programming Interfaces Explained - AWS," Amazon Web Services, Inc. Accessed: Feb. 21, 2023. [Online]. Available: https://aws.amazon.com/what-is/api/

[43] "About Twitter's APIs." Accessed: Feb. 21, 2023. [Online]. Available: https://help.twitter.com/en/rules-and-policies/twitter-api

[44] "Google Maps Platform," Google Developers. Accessed: Feb. 21, 2023. [Online].

Available: https://developers.google.com/maps
[45] H. Darabi, F. Karim, S. Harford, E. Douzali, and P. Nelson, “Detecting Current Job Market Skills and Requirements Through Text Mining,” in *2018 ASEE Annual Conference & Exposition Proceedings*, Salt Lake City, Utah: ASEE Conferences, Jun. 2018, p. 30284. doi: 10.18260/1-2--30284.
[46] C. Dhekne and S. Bansal, “MOOClink: An aggregator for MOOC offerings from various providers,” *J. Eng. Educ. Transform.*, vol. 2018, no. Special Issue, Jan. 2018, doi: 10.16920/jeet/2018/v0i0/120907.
[47] “What is Text Analysis? - Text Analysis and Mining Explained - AWS,” Amazon Web Services, Inc. Accessed: Jan. 04, 2024. [Online]. Available: https://aws.amazon.com/what-is/text-analysis/
[48] R. Egger and E. Gokce, “Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data,” in *Applied Data Science in Tourism*, R. Egger, Ed., in Tourism on the Verge. , Cham: Springer International Publishing, 2022, pp. 307–334. doi: 10.1007/978-3-030-88389-8_15.
[49] “NLTK :: nltk package.” Accessed: Jan. 04, 2024. [Online]. Available: https://www.nltk.org/api/nltk.html
[50] “What are Stop Words? A Guide to Stop Words (with List),” Semrush Blog. Accessed: Apr. 26, 2023. [Online]. Available: https://www.semrush.com/blog/seo-stop-words/
[51] D. Ladani and N. Desai, “Stopword Identification and Removal Techniques on TC and IR applications: A Survey,” Mar. 2020, pp. 466–472. doi: 10.1109/ICACCS48705.2020.9074166.
[52] G. L. Team, “Tokenising into Words and Sentences | What is Tokenization and it’s Definition?,” Great Learning Blog: Free Resources what Matters to shape your Career! Accessed: Apr. 26, 2023. [Online]. Available: https://www.mygreatlearning.com/blog/tokenization/
[53] D. M. Blei, “Latent Dirichlet Allocation”.
[54] Y. Pang, X. Xue, and A. S. Namin, “Predicting Vulnerable Software Components through N-Gram Analysis and Statistical Feature Selection,” in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2015, pp. 543–548. doi: 10.1109/ICMLA.2015.99.
[55] M. Allahyari *et al.*, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.” arXiv, Jul. 28, 2017. Accessed: Apr. 26, 2023. [Online]. Available: http://arxiv.org/abs/1707.02919
[56] S. H. Mohammed and S. Al-augby, “LSA & LDA topic modeling classification: comparison study on e-books,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 1, p. 353, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp353-362.
[57] S. Lunn, J. Zhu, and M. Ross, “Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice,” in *2020 IEEE Frontiers in Education Conference (FIE)*, Uppsala, Sweden: IEEE, Oct. 2020, pp. 1–9. doi: 10.1109/FIE44824.2020.9274270.
[58] “PRAW: The Python Reddit API Wrapper.” Python Reddit API Wrapper Development, Apr. 17, 2023. Accessed: Apr. 17, 2023. [Online]. Available: https://github.com/praw-dev/praw
[59] J. Kastrenakes, “Reddit reveals daily active user count for the first time: 52 million,” The Verge. Accessed: Apr. 26, 2023. [Online]. Available: https://www.theverge.com/2020/12/1/21754984/reddit-dau-daily-users-revealed

[60] S. Robot, “How Web Scraping Reddit Can Benefit You,” Scraping Robot. Accessed: Apr. 26, 2023. [Online]. Available: https://scrapingrobot.com/blog/web-scraping-reddit/
[61] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, “Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics,” *Soc. Media Soc.*, vol. 7, no. 2, Art. no. 2, Apr. 2021, doi: 10.1177/20563051211019004.
[62] B. Jeong, J. Yoon, and J.-M. Lee, “Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis,” *Int. J. Inf. Manag.*, vol. 48, pp. 280–290, Oct. 2019, doi: 10.1016/j.ijinfomgt.2017.09.009.
[63] juliancj, “What exactly is a ‘Sub’ reddit?,” r/help. Accessed: Mar. 30, 2023. [Online]. Available: www.reddit.com/r/help/comments/9f642j/what_exactly_is_a_sub_reddit/
[64] “csv — CSV File Reading and Writing,” Python documentation. Accessed: Apr. 17, 2023. [Online]. Available: https://docs.python.org/3/library/csv.html
[65] *Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2023. All rights reserved. Draft of February 3, 2024.*
[66] S. Kapadia, “Evaluate Topic Models: Latent Dirichlet Allocation (LDA),” Medium. Accessed: Apr. 26, 2023. [Online]. Available: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
[67] N. Kardam, S. Misra, and D. Wilson, “Is Natural Language Processing Effective in Education Research? A case study in student perceptions of TA support,” presented at the 2023 ASEE Annual Conference & Exposition, Jun. 2023. Accessed: Feb. 08, 2024. [Online]. Available: https://peer.asee.org/is-natural-language-processing-effective-in-education-research-a-case-study-in-student-perceptions-of-ta-support
[68] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.
[69] K. A. Severson *et al.*, “Data-driven prediction of battery cycle life before capacity degradation,” *Nat. Energy*, vol. 4, no. 5, Art. no. 5, May 2019, doi: 10.1038/s41560-019-0356-8.
[70] Y. Liu, R. Zhang, J. Wang, and Y. Wang, “Current and future lithium-ion battery manufacturing,” *iScience*, vol. 24, no. 4, Art. no. 4, Apr. 2021, doi: 10.1016/j.isci.2021.102332.
[71] A. Mosavi, P. Ozturk, and K. Chau, “Flood Prediction Using Machine Learning Models: Literature Review,” *Water*, vol. 10, no. 11, Art. no. 11, Nov. 2018, doi: 10.3390/w10111536.
[72] J. J. Duderstadt, “Engineering for a changing road, a roadmap to the future of engineering practice, research, and education,” 2007.
[73] D. Oguz and K. Oguz, “Perspectives on the Gap Between the Software Industry and the Software Engineering Education,” *IEEE Access*, vol. 7, pp. 117527–117543, 2019, doi: 10.1109/ACCESS.2019.2936660.
[74] C. Mabey, “Managing Graduate Entry,” *J. Gen. Manag.*, vol. 10, no. 2, Art. no. 2, Dec. 1984, doi: 10.1177/030630708401000205.
[75] H. Jang, “Identifying 21st Century STEM Competencies Using Workplace Data,” *J. Sci. Educ. Technol.*, vol. 25, no. 2, Art. no. 2, Apr. 2016, doi: 10.1007/s10956-015-9593-1.
[76] A. N. Azmi, Y. Kamin, M. K. Noordin, and A. N. Nasir, “Towards Industrial Revolution 4.0: Employers’ Expectations on Fresh Engineering Graduates,” *Int. J. Eng.*.
[77] M. Khorasani, “Can I transition from engineering to data science?,” Medium. Accessed:

Apr. 29, 2023. [Online]. Available: https://towardsdatascience.com/can-i-transition-from-engineering-to-data-science-2b55f6cdbb0

[78] H. J. Passow, "Which ABET Competencies Do Engineering Graduates Find Most Important in their Work?," *J. Eng. Educ.*, vol. 101, no. 1, Art. no. 1, 2012, doi: 10.1002/j.2168-9830.2012.tb00043.x.

[79] S. Male and R. King, "Improving Industry Engagement in Engineering Degrees," *N. Z.*, 2014.

[80] T. D. Fantz, T. J. Siller, and M. A. Demiranda, "Pre‐Collegiate Factors Influencing the Self‐Efficacy of Engineering Students," *J. Eng. Educ.*, vol. 100, no. 3, pp. 604–623, Jul. 2011, doi: 10.1002/j.2168-9830.2011.tb00028.x.

[81] V. Barr and C. Stephenson, "Bringing computational thinking to K-12: what is Involved and what is the role of the computer science education community?," *ACM Inroads*, vol. 2, no. 1, Art. no. 1, Feb. 2011, doi: 10.1145/1929887.1929905.

[82] K. J. Reid, J. O. Ladeji-Osias, C. Beauchamp, M. Dalal, T. Griesinger, and W. E. Eagle, "Design by Thread: The E4USA Engineering for Us All Curriculum," in *2020 IEEE Frontiers in Education Conference (FIE)*, Uppsala, Sweden: IEEE, Oct. 2020, pp. 1–6. doi: 10.1109/FIE44824.2020.9274008.

[83] J. Parilla and S. Liu, "TALENT-DRIVEN ECONOMIC DEVELOPMENT".

[84] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, "E-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics," *IEEE Access*, vol. 6, pp. 39117–39138, 2018, doi: 10.1109/ACCESS.2018.2851790.

[85] "Kaggle: Your Machine Learning and Data Science Community." Accessed: Feb. 04, 2024. [Online]. Available: https://www.kaggle.com/

[86] "Coursera | Degrees, Certificates, & Free Online Courses," Coursera. Accessed: Feb. 04, 2024. [Online]. Available: https://www.coursera.org/

[87] "Build new skills. Advance your career.," edX. Accessed: Feb. 04, 2024. [Online]. Available: https://www.edx.org