The Stated and Hidden Expectations - Applying Natural Language Processing Techniques to Understand Postdoc Experiences through Job Postings

Abstract

This paper represents a recent work applying natural language processing (NLP) techniques to generate insights on postdoc experiences from the job postings in engineering and computer science (CS). Postdoctoral positions are one of the important components of the academic career pipeline. It offers significant educational and professional opportunities, however, postdocs remain forgotten in the education community, especially in the field of engineering and CS with significant gender disparities in postdoc and faculty positions. In this work, we explore the NLP techniques to analyze the job postings for recruiting engineering and CS postdocs in the U.S. We utilized the Knowledge, Skills, and Attributes (KSAs) framework to characterize the KSAs as the stated expectations noted in the job postings. By applying the lexicon-based gender coding method on the job postings, we revealed that the majority of the postdoc job postings tend to use gendered language, which only further reinforces gender disparities in engineering and CS. Our findings indicated that it is important to provide clearly stated KSAs and gender neutral languages in the job postings to encourage underrepresented populations' participation and to support a sustainable academic career development for postdocs in engineering and CS.

1 Introduction

Postdoctoral positions (colloquially referred to as a "postdoc") are a common avenue for engineering and science students to prepare for competitive academic careers in STEM fields upon completion of their Ph.D. [1–3]. Effective postdoctoral experiences are facilitated through mentorship to help scholars develop deeper competencies and skills for the professoriate [4, 5]. Postdocs gain significant educational and professional experiences, however, they remain forgotten in the engineering and computer science (CS) education community, with few rigorous studies regarding postdocs in prominent journals. A major component of the postdoc experience is finding a faculty member or research group that will offer mentorship. Moreover, women and people of color tend to be underrepresented as faculty in engineering and computer science (CS), resulting in potential difficulties for graduates to find the same- and/or cross-gender and race mentors.

Presently, there is little known about how impressions of a postdoc position or certain mentors may influence recruiting experiences of postdocs. The purpose of this paper is to explore and to assess what are the stated and hidden expectations about postdocs during the recruitment process, and to examine the prevalence of gendered language in job postings which may influence aspiring postdocs. We applied Knowledge, Skills, and Attributes (KSAs) as the theoretical framework to answer the following research questions:

1. *What are the stated expectations for knowledge, skills, and attributes in the postdoc job postings required for engineering and CS postdocs?*
2. *Are there hidden expectations for gender perceptions from postdoc job postings?*

In this work, we answered these questions using natural language processing (NLP) techniques with Python. We collected data for engineering and CS postdoc job postings from publicly available web pages. Job postings are a rich source of information that can be used to extract relevant information about the required knowledge and skills in a particular area [6], such as project management [7], marketing [8], and big data[9], and so on. Using text mining to analyze the job postings to develop the job profiles used for recruitment has been effective and efficient [10]. It can also help to identify merging potential occupations [11] and to improve the quality of job matching [12]. Text mining is one of the major tasks of NLP [13], which has been a topic of interest in various educational research including e-learning [14], gamification in education [15], higher education [16, 17], STEM education [18–20] and more. A prior study exemplified how applying NLP on job postings can generate job market trends that offered additional educational consideration for CS education [21]. By utilizing content analysis on this textual data, this paper presented the stated expectations for postdocs by identifying the most frequently required knowledge, skills, and attributes from job postings. By applying lexicon-based gender-coding methods on the postdoc job postings, we were able to categorize masculine-coded, feminine-coded, and neutral job postings. As the advisors or PIs are the key contacts of the postdoc experiences, we argued the usage of gendered language in the job postings are indicators of the hidden expectations in the postdoc hiring process.

2 Literature Review

Postdoc positions aim to equip scholars for the professoriate with deeper methodological and content competency, writing, publishing ,and research skills, and research management skills [22–24]. A study by Andalib et al. [25] modeled the postdoc population as a labor force in a waiting queue. Using survey data from Ph.Ds. in the sciences and engineering, only 17% landed tenure-track and the average mean time in the queue was 2.9 years. Many new faculties in science and engineering, especially for women and minorities in the field, have at least one postdoctoral experience [26]. The study by Herschberg et al. [27] revealed a hasty and informal manner throughout the recruitment process for postdocs. Given the significantly increased number of externally funded, project-based postdoc positions, this results in the postdoc only being considered with their value towards the projects, not their personal and professional development through their positions. This qualitative comparative multiple case study also reported that the selection criteria for postdocs are often shaped by project conditions. This short-term mindset in the postdoc hiring process has been described in another study done by Knaub et al. [28]. This study outlined the informal hiring process and highlighted ties to the community as essential. It also revealed that PIs consider the ideal postdocs based on their background and how well they fit in with the group. A recent study on STEM postdocs suggested to support the success of postdocs through skills-based training will benefit the stakeholders in the academic ecosystem [29]. However, none of the studies mentioned above directly addressed the postdoc hiring issues in engineering and computer science education and no knowledge, skills, and attributes have been characterized to support the postdoctoral formation in this discipline.

Characterizing the knowledge, skills, and attributes (or sometimes, “abilities”) that learners must have in a particular setting is one way to capture implicit and explicit expectations for various career paths or levels of expertise. The use of KSAs to define boundaries in these settings have previously been employed in engineering education research, though not applied to postdocs to date. For example, Cox et al. [30] investigated engineering Ph.Ds. as stewards of the engineering discipline in academia, industry, and engineers who have migrated between the two trajectories and characterized a collection of KSAs. These KSAs of professionals can then be mapped back onto the doctoral education process, providing an adaptive blueprint for students and advisors to set professional development goals for doctoral students. Other work by Ahn and Cox [31] employed the KSA framework to understand the expectations and requirements of graduate students and postdocs engaging in undergraduate research mentorship. Other applications of the KSA approach to competency development in engineering, whether or not they are labeled specifically as "KSAs," have been applied by studies such as those by Ngyuen [32], Harun et at. [33], and Rayner and Papakonstantinou [34]. In each of these studies, the aims of the resulting analyses and lists of attributes are to align educational outcomes with the needs of employers, whether that be in academia and industry. Walther and Radcliffe [35] posit that a more nuanced perspective on competency development is required that includes attitudes and self-image as a means of understanding professional competence.

Of note, the methods by which KSAs are generated determine their applicability and scope. For example, in Cox's work, the KSAs were generated after 40 intensive qualitative interviews with engineering Ph.Ds. from four different trajectories of PhD holders, and the resulting KSAs were characterized through rigorous qualitative analysis. Therefore, these KSAs yield both explicit expectations (competencies identified by the participants), and implicit expectations that resulted from inductive qualitative analysis from the interviews. Other methods for yielding competency lists are more prescribed, such as those from a document analysis, but yield more limited results. Additionally, there were prior works that have alluded to KSAs about postdoc career but have not been directly applied to postdoc recruitment. Davis' work [36] outlined teaching skills, proposal writing, and project management as skills associated with positive outcomes for postdocs experiences through an empirical approach. Nowell et al. [37] identified required skill sets for postdocs to pursue career development opportunities. Those skill sets were generated from a mixed-methods study using a cross-sectional survey and qualitative interviews. Layton et al. [38] conducted a survey study that captured skills through academic career courses specifically designed to increase postdocs career awareness and confidence. Other related works [39, 40] implied the practical skills that enhance postdoc career development including but not limited to mentorship, project management, communication skills, networking, and leadership skills. In the present study, we seek to explore a mixed approach to a predefined KSA dictionary based on prior literature and to mine postdoctoral job postings for mentions of knowledge, skills, and attributes required for a successful candidate.

Beyond KSAs, gender bias in language can also contribute to the expectations behind a given text. Gender bias emerges in language choices used in writing and verbal communication [41]. Several studies have demonstrated that gender bias is always hidden in job postings that can discourage certain applicants from applying [42–46]. Gaucher et al. showed job postings for male-dominated areas employed greater masculine wording than those within female-dominated areas [42]. The study also indicated that job postings with more masculine than feminine wording, had participants that perceived more men within these occupations, which often makes the job less appealing to women. The perceptions of belongingness were also studied in this work and results confirmed that women anticipated less belongingness in jobs where the postings were masculine worded. Men were only slightly more likely to be interested in masculine worded jobs than those feminine worded and no effects were found on men's feelings of belongingness within the occupation. Tang et al. conducted a longitudinal analysis for gender bias in the job market in 2017 [47]. This study reported an increasing shift away from masculine-biased job postings over the years as employers use less gendered wording than they did 10 years ago. The prior works related to gender wording in job postings focused on the impacts and implications for the overall job markets. It has rarely explored the gendered wording usage on any specific gender dominant disciplines, such as engineering and computer sciences.

Considering the postdoctoral stage of education as an "intermediate" professional stage of development is quite interesting when considering a KSA-based approach to understanding competencies. While postdocs have earned Ph.Ds. and are capable of fully-independent research, they are expected to carry both administrative and research management burdens for their advising professor (under the aim of learning to manage a research group), and acquire new skills or hone existing skills. Because the postdoctoral education literature has not characterized the educational development of postdocs, it is difficult through literature to discern which KSAs are expected of postdocs as they are incoming to postdoctoral positions, and which KSAs they are expected to learn through their time as a postdoc. Therefore, an essential first step is to gather information on what explicit skills are requested of postdocs in position postings. Meanwhile, women postdocs in engineering and CS remain underrepresented. The most recent data from the 2018 National Center for Science and Engineering Statistics (NCSES) Survey reported only 23.6% women for postdoc positions in engineering and 20.6% women postdocs in computer sciences [48]. By extending the prior works about gendered languages in postdoc job postings to the male-dominated engineering and CS discipline, our work seeks to explore the implicit expectations for postdocs in this domain to further our understanding.

3 Methods

3.1 Research Design

The goal for this study is to understand the expectations for postdocs through the job postings. In our study, we attempted to combine job posting analysis and NLP to characterize the postdocs profiles that may help us better understand the initial hurdles in engineering and computer science academic career development via the postdoc position. The overall research process includes four major components. First, we created the dataset by a combination of automatic and manual web scraping. Web scraping is the process of extracting unstructured data from web pages that can be used to build datasets with structured data [49]. Common methods for web scraping vary from manually copying and pasting data from a page to automatically accessing the page to obtain useful information through programming languages [50]. Second, we processed the collected raw text data to transform the textual data into usable forms to conduct data analysis [51], which is preliminary work necessary to perform any NLP tasks [52]. Third, we performed textual analysis on the job postings to generate knowledge and insights on KSAs and the usage of gendered language. Lastly, we presented the analysis results through visualization. Python 3.7.6 has been used to perform all the tasks of this research process.

3.2 Data Collection and Processing

We collected the postdoc job postings through scraping Indeed.com and university career pages. We used Python BeautifulSoup [53] to scrape the job searching website with "engineering postdoc" and "computer science postdoc" as the search keywords. To collect data from university career pages, we came up with a list of 120 universities that consist of a combination of the top 25 universities in research expenditures in engineering and top universities with most graduate students in engineering, both based on data from 2018. We only included job postings if the postings were for postdocs or equivalent positions and "engineering" or "computer science" were identified as keywords in the hiring department, required skills, or as the required Ph.D. discipline. We limited the geographic regions of the data collection to the U.S to create the representative sample for this study. After removing the duplicate records, our final dataset included n = 823 job postings for engineering and computer science postdocs.

In order to capture the KSAs and the gendered language used in the job postings, we manually developed three dictionaries used in the data analysis. We first explored using the KSAs framework to develop the dictionary. KSAs mentioned in related literature on postdoc studies [36–40] were mapped into 13 KSAs features to form this dictionary (Table 1). During this process, we realized that most postdoc job postings were so vague that at times, they simply listed the domain discipline as an indicator for skills or experiences required (e.g., bioengineering). The less vague job postings described technical skills required for completing a specific research project under a domain discipline (e.g., proficiency in microcontroller design and programming for electronic engineering research) or it simply listed experiences or background requirements for specific labs or research domains (e.g., experience in marsh ecology to support modeling sea-level rise). Due to the unstructured nature of the job postings, it was difficult to access the specific skills needed for the positions, therefore we adopted a disciplinary approach to classify the technical skills. We developed a domain related dictionary by using the Classification of Instructional Programs (CIP) codes for engineering and computer sciences. A list of all 78 values identified for this study is provided in Appendix A. We used CIP to capture the domain disciplines as an indicator for domain specific technical skills associated with the research projects when hiring for postdocs. To prepare for identifying the usage of gendered language in the job postings, we utilized the word lists developed by Gaucher et al. [42]. All the text data from job postings and the dictionaries have been converted into lowercase and removed the punctuation, special characters, and stopwords. We used Python Pandas [54], NumPy[55], and NLTK [56] libraries to complete this data processing required for the NLP techniques.

The processed dataset with structured features is shown in Table 2. We first categorized the job postings based on the types of institutions. Postdoc appointments under universities were assigned to "academia." Other appointments at national laboratories, industry research centers, or corporations were categorized as "non-academia." To further extract the structure from the text data, the KSAs and domain discipline dictionaries were applied to analyze the job posting data. The word frequencies were calculated based on the two dictionaries. Two lists of identified KSAs and domain discipline were generated and appended to the dataset. The gendered word list has also been applied to the job postings data using the tool developed based on the same studies, called gender decoder [57]. Each job posting was categorized as feminine-coded, masculine-coded, or neutral based on the calculated relative proportions of the gendered words. Another list with identified gendered language categories was also added to the dataset.

3.3 Data Analysis and Visualization

In order to capture the overall themes of the job postings, we first plotted a WordCloud visualization on the full text of the collected job ads. WordCloud is a data visualization technique that illustrates

Table 1: KSAs Features Dictionary

KSAs Features	Examples of KSAs Features
Academic Career Skills	- Grants/awards adjudication - Mock interviews - Writing research and teaching statements - Identify career pathways - Goal setting
Academic Writing	- Writing research publications including journals, papers, technical documents, etc. - Writing and reviewing grants
Career Planning	- Identify careers that match goals - Prepare for job applications - Postdocs job fairs/workshops
Communication Skills	- Attending conferences/seminars/workshops - Delivering research presentations - Interacting and collaborating with other researchers - Preparing job talks - Excellent written and spoken English
Industry Career Skills	- Networking with industry - Resume and cover letter writing - Transitioning from postdoc to industry
Leadership Skills	- Leading a collaborative research team in the lab - Leadership on research projects - Diversity awareness - Openness to critique
Mentoring	- Mentoring graduate students and junior postdocs - Managing small groups - Peer - mentorship - Access to role models
Networking	- Networking among postdocs - Identifying collaborators
Personal Reflection	- Identifying professional interests and values
Project Management	- Project assignments allocation - Project financial management, funding allocation - Not just doing, but finish projects and publications
Teaching and Learning	- Giving guest lectures in classes - Teaching a course - Developing teaching philosophy/teaching dossier
Time Management	- Managing deliverables to meet the deadline - Ability to work under time pressure

Table 2: Data Features and Descriptions

Feature	**Description of Feature**	**Example(s)**
full_text	The complete job description for a postdoc position	The successful postdoc candidate will join a team that combines computational and experimental researchers using computational systems. Applicants must have a doctoral degree in biomedical engineering, chemical engineering, or equivalent with a demonstrated record of innovative scientific accomplishments as evidenced by papers published or accepted in premier journals. Qualified candidates must demonstrate outstanding communication skills, have a strong commitment to science, and work well within a group.
institution_type	Binary values to indicate types of the hiring institution.	“0” = non academic institutions “1” = academic institutions
ksa	The identified KSAs from the full job descriptions after data processing	[’academic writing’, ’communication skills’, ’leadership skills’, ’project management’]; [’communication skills’, ’leadership skills’, ’time management’]
domain	The identified domain discipline from the full job descriptions after data processing	[’computer science’, ’artificial intelligence’, ’software engineering’]; [’biochemistry engineering’, ’biomedical engineering’]

gender_coding	The identified gender languages categories from the full job descriptions after data processing	masculine-coded;
		neutral;
		feminine-coded

text data based on its frequency in the text. The frequency of each word corresponds to the size of words shown in the cloud [58]. Next, frequency counts for KSAs and domain discipline were calculated, respectively. Descriptive data analysis was applied to summarizing the corresponding skills stated in job postings for postdocs in engineering and computer science. The descriptive statistics of gendered language categories were calculated to characterize the hidden expectations. To investigate whether there were any significant differences between data features, we explored Kruskal Wallis tests on the dataset to further our understanding. Additional Python libraries applied to the analysis and visualization including Matplotlib [59] and Scipy [60].

4 Results

4.1 Assessing postdoc job postings

After removal of stopwords and special characters in the job postings, we utilized a word cloud to illustrate the most frequently used terms in the overall job postings. As mentioned in the method section, the higher the frequency of the words appeared in the corpus, the larger size will be displayed in the word cloud. As shown in Figure 1, "research," "project," "experience," "lab," "system," and "data" are the high frequency words mentioned most in all of the job postings. Some other obvious words in the word cloud related to the KSAs or domain discipline included "computer sciences," "modeling," "communication skills," "machine learning," and "biomedical engineering." The domain discipline codes refer to technical skills usually associated with said discipline. These codes were included in order to capture the technical skills requested in the postings, but also the vague nature of how those skills are communicated. In this case, "biomedical engineering" is a popular domain discipline because labs and departments under medical schools have a higher demand for hiring engineering postdocs and require the skills commonly associated with this discipline.

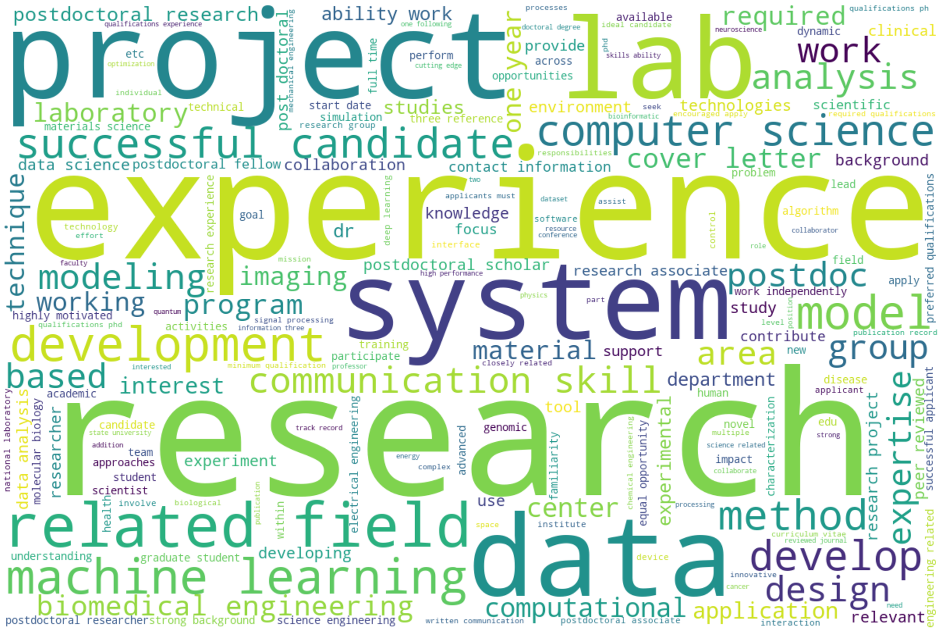

Figure 1: Word Cloud of Job Postings for Postdocs in Engineering and Computer Sciences

4.2 Stated expectations in job postings

The first stated expectation is the KSAs identified from the postdoc job postings. The top 5 stated KSAs in the job postings are "communication skills" (69.9%), "academic writing" (61.4%), "leadership skills" (34.5%), "mentorship" (14.8%), and "teaching and learning" (7.9%), as shown in Figure 2. As for domain discipline required for postdoc positions, Figure 3 demonstrated the top 10 domain disciplines mentioned in our dataset included "General engineering" (71.8%) and "computer sciences" (33.2%) among the top 3 domains, which is not a surprise as we limited the data collection in the engineering and computer science domains. More than half of the postings identified "statistics" (53.5%) in the postings which makes it a top 2 domain discipline.

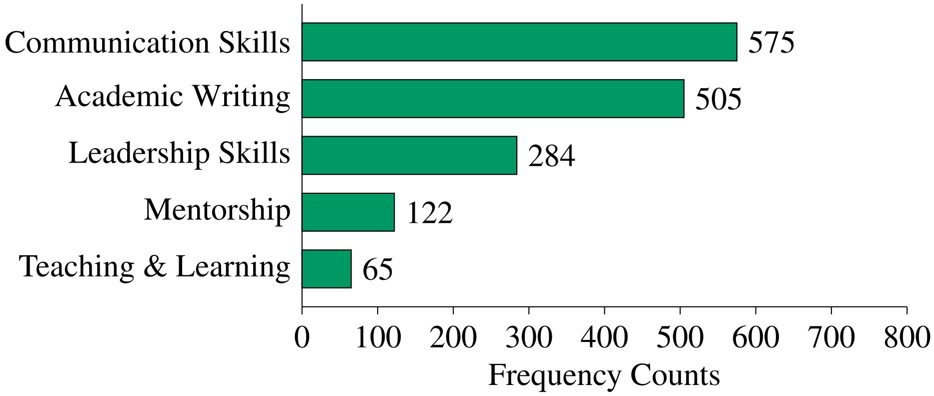

Figure 2: Top 5 identified KSAs in Engineering and Computer Sciences

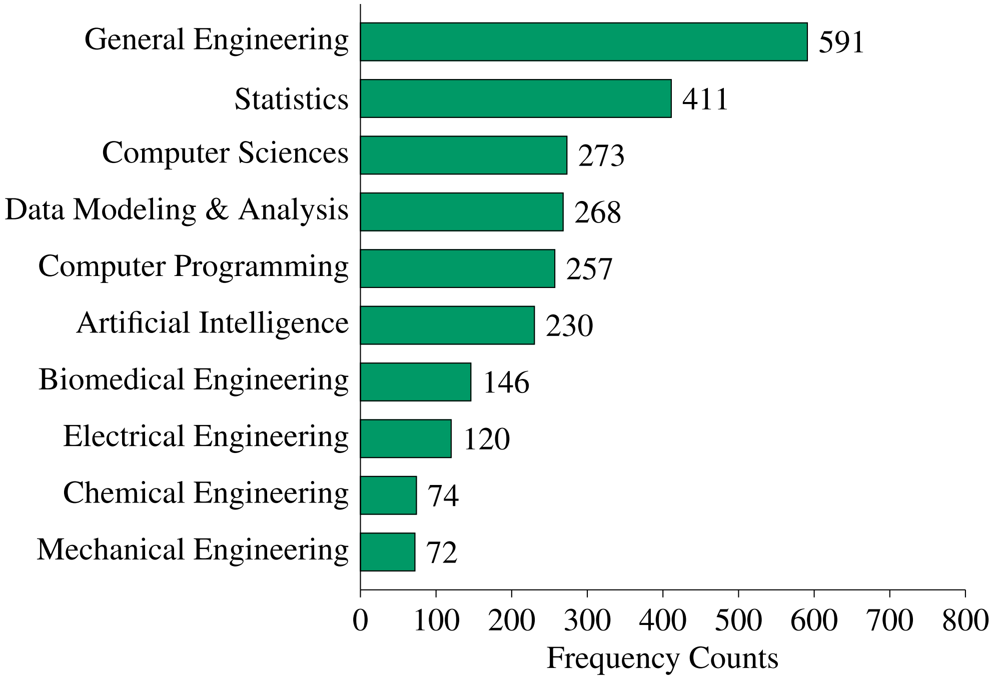

Figure 3: Top 10 Identified Domain Disciplines in Engineering and Computer Sciences

4.3 Hidden Expectations in job postings

A bar plot of the gendered language is provided in Figure 4. This plot grouped the categorized gendered language encoding into "masculine-coded," "feminine-coded," and "neutral" by the number of occurrences. More than 80% of our data demonstrated usage of gendered language in the postdoc job posting. A total of 42.9% of the job postings are masculine-coded, which is slightly higher than feminine-coded postings with 40.7%. Neutral job postings consisted of 17.4%. Each of the groups in Figure 4 have been divided into academic or non-academic job postings. For postdoc positions at academic institutions, 42.6% reported masculine-coded postings, 39.7% for feminine-coded and 17.7% neutral job postings. In the meantime, for job postings at non-academic institutions, the masculine, feminine and neutral coded job postings were 35.7%, 50.0% and 14.3%, respectively. For postdoc positions, depending on different types of hiring institutions, academic institutions tend to use more masculine-coded language in postings and non-academic institutions are more likely to be feminine-coded language.

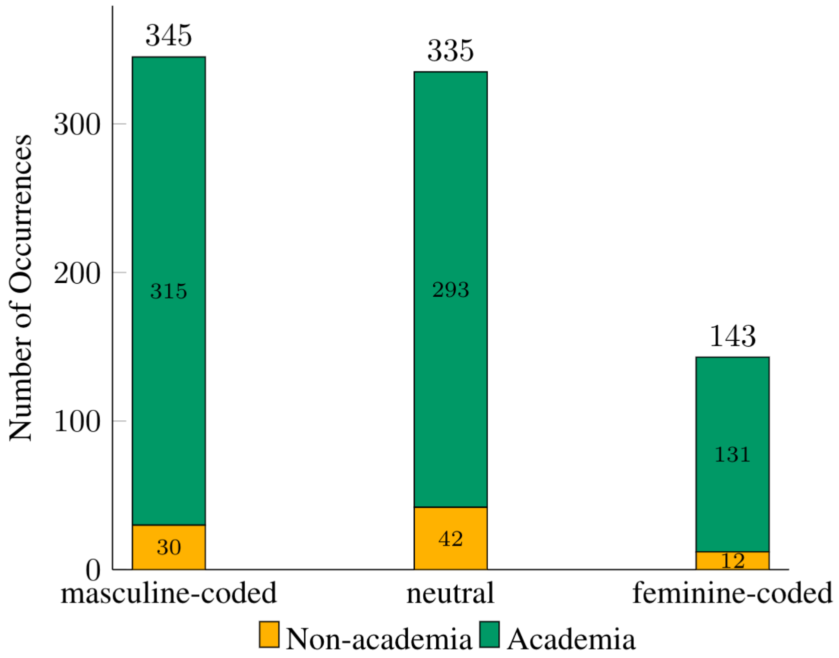

Figure 4: Gendered Language Decoding

4.4 Relationships between Data Features

We further performed Kruskal Wallis tests on relationships between the categorical data features. We accepted significant differences in results only if a p-value <0.05 was found in the tests. We first compared whether there were any significant differences between identified KSAs and domain discipline between two groups academic and non-academic institutions, respectively. Significant differences were found for KSAs and for domain discipline. For non-academic postdoc positions, it tends to list more KSAs and domain discipline in the job postings. Then, the gendered language was found to be significantly different between academic and non-academic institutions postdoc job postings. This further supported the descriptive results provided in section 4.3. Lastly, we tested whether identified KSAs and domain discipline are significantly different between the gendered language categories. Significant differences were found between identified KSAs and gendered language in job postings. The feminine-coded postings with a slightly higher likelihood to identify more KSAs. No such difference was found between gendered language and domain discipline in the job postings.

Table 3: Kruskal Wallis Test with Significant Results

Data Feature Tested		*P* - value (<0.05)
Academic vs. non academic	KSAs	$p = 0.001$
	Domain discipline	$p = 0.036$
	Gendered language	$p = 0.042$
Gendered language	KSAs	$p = 3.29 \text{ x } 10^{-5}$

5 Discussion

Our analysis suggested that the most prevalent KSAs identified through the postdocs job postings are communication skills and academic writing. Our study echoes the prior studies on the skills related to writing and publishing [22–24] and emphasizes the importance of communication skills in the academic pipeline for engineering and CS postdoc scholars. As indicated in prior literature, the majority of the postdoc positions are shaped around specific projects [27], our results are consistent with prior studies as the word cloud identified the high frequency of words used around the projects. To better capture the project-related skills, we further extended the KSAs framework with a set of domain disciplines based on the CIP codes as an indicator for the associated technical skills. Domain disciplines like "general engineering" and "computer science" are intuitive, since we purposefully sought out engineering and computer science job postings. But other technical skills like "statistics" and "data modeling and analysis" were shown at the top of the identified domain discipline, followed by "computer programming" and "artificial intelligence." Our results further suggested that postdoc positions expect a high level of computational proficiency from data analysis, statistical modeling, programming, and beyond. Other specific technical skills required of postdocs are difficult to access from the job postings due to the vague boundary between discipline and skills being communicated. When an entire discipline, like "biomedical engineering", represents the skill set required, the applicant has no idea which specific skills are actually necessary for the job and may assume they must be a "master of the discipline."

Regarding the hidden expectations demonstrated in postdoc postings, the results demonstrate that the majority of the postdoc job postings tend to use gendered language, which may further reinforce gender disparities in engineering and CS postdocs appointments. The percentage of the masculine-coded job postings is only slightly higher than the feminine-coded postings. One of the reasons is, as indicated in prior work [47], that an increasing shift away from masculine-biased job postings over the year. On the other hand, through the data collection process, we realized that there are many universities' career websites utilizing Workday, a third-party application for talent recruitment. This third-party application has partnered with Textio that integrates the data-driven language insights for recruiters and hiring managers when they write job posts in Workday [61]. Textio is an online service based on Gaucher et al. encoded list that helps to minimize the gender bias in job postings [47]. We argue that those job postings published through Workday empowered university recruitment sites have been gender neutralized. Moreover, postdoc postings from non-academic institutions reported less masculine-coded, which may encourage more female applicants for postdoc careers outside of academia. Furthermore, the feminine-coded postings with a slightly higher likelihood to identify more KSAs. As our study adopted the lexicon approach for detecting the gendered language, the KSAs identified through the framework included skills or attributes that are perceived as feminine language in the word list from a prior study [42]. Examples for those KSAs include communication, interpersonal, and collaboration and so on. Gaucher's work [42] suggested that, unlike women, there are limited impacts for men about feminine-worded job postings regarding their perceptions on job appeal and perceived belongingness for the occupations. We encourage postdoc job postings to increase more KSAs that may help improve women's perceived postdoc job appealingness and belongingness.

Another hidden expectation revealed during the data collection process is the necessity to have connections in the research community. The fact that most job postings were so vague (only stating the discipline) or requiring job applicants to contact the PI for position details directly contributes to the pervasiveness of informal hiring processes for postdocs. We discovered that there is a significant amount of postdoc opportunities that were listed on the department or lab websites with no further details, simply stating "Postdoctoral opportunity." PIs develop mentoring plans as a way to eliminate the uncertainty of the postdoc requirements in this hiring process. But the mentoring plan is shared with potential postdocs who have established ties with the PIs. Thus, connections to the people in the research community becomes the key for applicants seeking postdoc opportunities, which supported the findings in a prior study [28]. We also know, from prior work, wording and requirements listed in the job postings influences women's willingness to apply [45, 46]. If for example, women will not apply for jobs that they can not claim full competency in, then how do they react to a job posting with zero qualifications stated [62]? We aimed to apply the KSAs framework to characterize the skills for postdocs in engineering and CS; however, the current analysis could only be based on those postdoc positions that provided a job description for recruiting. For those postdocs through the informal hiring process, the KSA framework is no longer applicable to this scenario. The application process for such postdoc positions starts with reaching out to PIs directly. PIs that recruit from this informal process expect postdocs to come to them directly for any potential opportunities, which may alienate potential candidates that are not familiar with the academic hiring process.

6 Limitations and Future Work

Although taking a computational approach on the text data demonstrated to be efficient and effective, it should not replace the human researcher's role. The KSAs identified through our approach are likely to be more limited than deeper dives mining other documents, generating lists of KSAs from a combination of literature, or interviewing postdoctoral scholars and faculty mentors. Moreover, manual validation is often required and we seek to validate the findings in future studies by interviewing postdocs as well as their PIs. We will use this study as a pilot to decide the scope of the interview questions and to develop interview protocols. Furthermore, our data is limited in size which led our work to adopt the lexicon-based approach and set constraints for us to explore deep learning approaches. Recent work has proposed a deep learning approach with improved accuracy to identify the gender stereotypes in text data [63]. We would like to expand the data size and investigate advanced NLP techniques with deep learning models. In addition, the majority of the current analysis is on the postdoc positions with published job postings. As the informal hiring consists of one of the main recruitment channels for engineering and CS postdocs, we intend to incorporate studies about their experiences through the informal hiring and extend the theoretical framework to social capital theory to better guide our studies in the future.

7 Conclusion

In this work, we demonstrated the ability to apply NLP techniques to generate meaningful insights on postdoctoral experiences in engineering and CS. By analyzing the job postings through the computational approach, our results characterized the KSAs profiles and disclosed the tendency to use masculine coded language. We suggest PIs to scrutinize the job postings to avoid any vague statements and to provide potential applicants clearly stated expectations for postdoc positions. To create a more sustainable environment for postdocs, it is noteworthy for academia to clearly provide resources that support the development of communication skills, academic writing, and computational proficiency. Furthermore, to broaden participation in the academic pipeline for postdocs, it is important to reconsider the language used in such postings. This work is also intended to call attention to the gendered language and how it could potentially discourage underrepresented populations and to provide a foundation for further exploring search, screen and hiring in engineering and CS postdocs.

References

[1] M. Nerad and J. Cerny, "Postdoctoral patterns, career advancement, and problems," *science*, vol. 285, no. 5433, p. 1533—1535, 1999.

[2] G. S. Åkerlind*, "Postdoctoral researchers: roles, functions and career prospects," *Higher Education Research & Development*, vol. 24, no. 1, p. 21—40, 2005.

[3] G. S. Åkerlind, "Postdoctoral research positions as preparation for an academic career," *International Journal for Researcher Development*, 2009.

[4] A. K. Scaffidi and J. E. Berman, "A positive postdoctoral experience is related to quality supervision and career mentoring, collaborations, networking and a nurturing research environment," *High. Educ*, vol. 62, no. 6, p. 685–698, 2011.

[5] J. M. Faupel–Badger, K. Raue, D. E. Nelson, and S. Tsakraklides, "Alumni perspectives on career preparation during a postdoctoral training program: A qualitative study," *CBE—Life Sciences Education*, vol. 14, no. 1, p. ar1, 2015.

[6] M. Pejic–Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Text mining of industry 4.0 job advertisements," *International journal of information management*, vol. 50, p. 416—431, 2020.

[7] K. Ahsan, M. Ho, and S. Khan, "Recruiting project managers: A comparative analysis of competencies and recruitment signals from job advertisements," *Project Management Journal*, vol. 44, no. 5, p. 36–54, 2013.

[8] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on big data in marketing: A text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, p. 1–7, 2018.

[9] A. Gardiner, C. Aasheim, P. Rutner, and S. Williams, "Skill requirements in big data: A content analysis of job advertisements," *Journal of Computer Information Systems*, vol. 58, no. 4, p. 374–384, 2018.

[10] G. K. Palshikar, R. Srivastava, S. Pawar, S. Hingmire, A. Jain, S. Chourasia, and M. Shah, "Analytics–led talent acquisition for improving efficiency and effectiveness," in *Advances in analytics and applications*. Springer, 2019, p. 141—160.

[11] M. Mezzanzanica, "Italian web job vacancies for marketing–related professions. symphonya," *Emerging Issues in Management, (*, vol. 3, p. 110–124, 2017.

[12] Y. Kino, H. Kuroki, T. Machida, N. Furuya, and K. Takano, "Text analysis for job matching quality improvement," *Procedia computer science*, vol. 112, p. 1523–1530, 2017.

[13] A. Kao and S. R. Poteet, *Natural language processing and text mining*. Science Business Media: Springer, 2007.

[14] J. Hung, “Trends of e–learning research from 2000 to 2008: Use of text mining and bibliometrics,” *British Journal of Educational Technology*, vol. 43, p. 5–16, 2012. [Online]. Available: https://doi.org/10.1111/j.1467–8535.2010.01144.x

[15] J. Martí–Parreño, E. Méndez–Ibáñez, and A. Alonso–Arroyo, “The use of gamification in education: a bibliometric and text mining analysis,” *Journal of computer assisted learning*, vol. 32, no. 6, p. 663—676, 2016.

[16] C. L. Santos, P. Rita, and J. Guerreiro, “Improving international attractiveness of higher education institutions based on text mining and sentiment analysis,” *International Journal of Educational Management*, 2018.

[17] A. Karami, C. N. White, K. Ford, S. Swan, and M. Y. Spinel, “Unwanted advances in higher education: Uncovering sexual harassment experiences in academia with text mining,” *Information Processing & Management*, vol. 57, no. 2, p. 102167, 2020.

[18] M. Jin and H. K. Ko, “Analysis of trends in mathematics education research using text mining,” *Communications of Mathematical Education*, vol. 33, no. 3, p. 275–294, 2019.

[19] S. Chopra, H. Gautreau, A. Khan, M. Mirsafian, and L. Golab, “Gender differences in undergraduate engineering applicants: A text mining approach.” *International Educational Data Mining Society*, 2018.

[20] A. Fortino, Q. Zhong, W. C. Huang, and R. Lowrance, “Application of text data mining to stem curriculum selection and development,” in *2019 IEEE Integrated STEM Education Conference (ISEC)*. IEEE, 2019, p. 354—361.

[21] 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)*. IEEE, 2020, p. 1—9.

[22] B. Rybarczyk, L. Lerea, P. K. Lund, D. Whittington, and L. Dykstra, “Postdoctoral training aligned with the academic professoriate,” *Bioscience*, vol. 61, no. 9, p. 699–705, 2011.

[23] X. Su, “Postdoctoral training, departmental prestige and scientists’ research productivity,” *J. Technol. Transf.*, vol. 36, no. 3, p. 275–291, 2011.

[24] Á. Borrego, M. Barrios, A. Villarroya, and C. Ollé, “Scientific output and impact of postdoctoral scientists: A gender perspective,” *Scientometrics*, vol. 83, no. 1, p. 93—101, 2010.

[25] M. A. Andalib, N. Ghaffarzadegan, and R. C. Larson, “The postdoc queue: A labour force in waiting,” *Systems Research and Behavioral Science*, vol. 35, no. 6, p. 675–686, 2018.

[26] W. M. Reichert, T. Daniels–Race, and E. H. Dowell, “Time–tested survival skills for a publish or perish environment,” *Journal of Engineering Education*, vol. 91, no. 1, p. 133–137, 2002.

[27] C. Herschberg, Y. Benschop, and M. Van den Brink, “Precarious postdocs: A comparative study on recruitment and selection of early–career researchers,” *Scandinavian Journal of Management*, vol. 34, no. 4, p. 303—310, 2018.

[28] A. V. Knaub, M. Jariwala, C. R. Henderson, and R. Khatri, “Experiences of postdocs and principal investigators in physics education research postdoc hiring,” *Physical Review Physics Education Research*, vol. 14, no. 1, p. 010152, 2018.

[29] M. Ålund, N. Emery, B. J. Jarrett, K. J. MacLeod, H. F. McCreery, N. Mamoozadeh, J. G. Phillips, J. Schossau, A. W. Thompson, A. R. Warwick *et al.*, “Academic ecosystems must evolve to support a sustainable postdoc workforce,” *Nature ecology & evolution*, vol. 4, no. 6, p. 777—781, 2020.

[30] M. F. Cox, T. Zephirin, N. Sambamurthy, B. Ahn, J. London, O. Cekic, A. Torres, and J. Zhu, “Curriculum vitae analyses of engineering ph. ds working in academia and industry,” *transformation*, vol. 15, p. 16, 2013.

[31] B. Ahn and M. F. Cox, “Knowledge, skills, and attributes of graduate student and postdoctoral mentors in undergraduate research settings,” *Journal of Engineering Education*, vol. 105, no. 4, p. 605—629, 2016.

[32] D. Q. Nguyen, “The essential skills and attributes of an engineer: A comparative study of academics, industry personnel and engineering students,” *Global J. of Engng. Educ*, vol. 2, no. 1, p. 65–75, 1998.

[33] H. Harun, R. Salleh, M. N. R. Baharom, and M. A. Memon, “Employability skills and attributes of engineering and technology graduates from employers’ perspective: Important vs. satisfaction,” *Global Business and Management Research*, vol. 9, no. 1s, p. 572, 2017.

[34] G. M. Rayner and T. Papakonstantinou, “Employer perspectives of the current and future value of stem graduate skills and attributes: An australian study,” *Journal of Teaching and Learning for Graduate Employability*, vol. 6, no. 1, p. 100–115, 2015.

[35] J. Walther and D. F. Radcliffe, “The competence dilemma in engineering education: Moving beyond simple graduate attribute mapping,” *Australasian Journal of Engineering Education*, vol. 13, no. 1, p. 41–51, 2007.

[36] G. Davis, “Improving the postdoctoral experience: An empirical approach,” in *Science and engineering careers in the United States: An analysis of markets and employment*. Chicago: University of Press, 2009, p. 99–127.

[37] L. Nowell, G. Ovie, N. Kenny, and M. Jacobsen, “Postdoctoral scholars’ perspectives about professional learning and development: a concurrent mixed–methods study,” *Palgrave Communications*, vol. 6, no. 1, p. 1–11, 2020.

[38] R. L. Layton, V. S. H. Solberg, A. E. Jahangir, J. D. Hall, C. A. Ponder, K. J. Micoli, and N. L. Vanderford, “Career planning courses increase career readiness of graduate and postdoctoral trainees,” *F1000Research*, vol. 9, 2020.

[39] M. B. Omary, Y. M. Shah, S. Schnell, S. Subramanian, M. S. Swanson, and M. X. O’Riordan, “Enhancing career development of postdoctoral trainees: act locally and beyond,” *The Journal of physiology*, vol. 597, no. 9, p. 2317, 2019.

[40] T. J. Didiano, L. Wilkinson, J. Turner, M. Franklin, J. H. Anderson, M. Bussmann, D. Reeve, and J. Audet, “I have a ph. d.! now what? a program to prepare engineering ph.

d.'s and postdoctoral fellows for diverse career options." American Society for Engineering Education, 2019.

[41] M. Menegatti and M. Rubini, "Gender bias and sexism in language," in *Oxford Research Encyclopedia of Communication*, 2017.

[42] D. Gaucher, J. Friesen, and A. C. Kay, "Evidence that gendered wording in job advertisements exists and sustains gender inequality." *Journal of personality and social psychology*, vol. 101, no. 1, p. 109, 2011.

[43] S. L. Bem and D. J. Bem, "Does sex–biased job advertising "aid and abet" sex discrimination?" *Journal of Applied Social Psychology*, vol. 3, no. 1, p. 1973, 1973.

[44] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson, "Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, p. 1914—1933.

[45] L. Wille and E. Derous, "When job ads turn you down: how requirements in job ads may stop instead of attract highly qualified women," *Sex Roles*, vol. 79, no. 7, p. 464—475, 2018.

[46] T. Hentschel, S. Braun, C. Peus, and D. Frey, "Sounds like a fit! wording in recruitment advertisements and recruiter gender affect women's pursuit of career development programs via anticipated belongingness," *Human Resource Management*, 2020.

[47] S. Tang, X. Zhang, J. Cryan, M. J. Metzger, H. Zheng, and B. Y. Zhao, "Gender bias in the job market: A longitudinal analysis," *Proceedings of the ACM on Human–Computer Interaction*, vol. 1, no. CSCW, p. 1—19, 2017.

[48] "Survey of graduate students and postdoctorates in science and engineering," 2018. [Online]. Available: https://www.nsf.gov/statistics/srvygradpostdoc/.

[49] D. S. Sirisuriya *et al.*, "A comparative study on web scraping," 2015.

[50] S. Munzert, C. Rubba, P. Meißner, and D. Nyhuis, *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons, 2014.

[51] Sciforce, "Text preprocessing for nlp and machine learning tasks," May 2020. [Online]. Available: https://medium.com/sciforce/text–preprocessing–for–nlp–and–machine–learning–tasks–3e077aa4946e

[52] R. Cheng, "Text preprocessing with nltk," Jun 2020. [Online]. Available: https://towardsdatascience.com/nlp–preprocessing–with–nltk–3c04ee00edc0

[53] L. Richardson, "Beautiful soup documentation," *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018]*, 2007.

[54] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 2010, p. 51—56.

[55] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau,

E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, “Array programming with numpy,” *Nature*, vol. 585, no. 7825, p. 357—362, 2020.

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

[57] K.Matfield and M.Nordlund, “lovedaybrooke/gender–decoder.” [Online]. Available: https://github.com/lovedaybrooke/gender–decoder

[58] “Wordcloud for python documentation.” [Online]. Available: https://amueller.github.io/word_cloud/

[59] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing in Science Engineering*, vol. 9, no. 3, p. 90–95, 2007.

[60] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, “Scipy 1.0: fundamental algorithms for scientific computing in python,” *Nature methods*, vol. 17, no. 3, p. 261—272, 2020.

[61] “Textio brings augmented writing to workday recruiting.” [Online]. Available: https://textio.com/blog/textio–brings–augmented–writing–to–workday–recruiting /31236002335

[62] T. S. Mohr, “Why women don’t apply for jobs unless they’re 100% qualified,” *Harvard Business Review*, vol. 8, 2014.

[63] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao, “Detecting gender stereotypes: Lexicon vs. supervised learning methods,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, p. 1—11.

Appendix A

List of CIP codes for engineering and CS to identify domain discipline

artificial intelligence
computer programming
statistics
computer sciences
computer software
information technology
informatics
information sciences
data processing technology
computer systems analysis
web & multimedia design
data modeling, analysis
& administration
computer graphics
computer modeling
& simulation
computer media applications
computer systems networking
telecommunications
network and system
administration
system management
networking management
computer and information
systems security
web & multimedia
management
information technology
management

general engineering
aerospace, aeronautical
& astronautical engineering
agricultural engineering
biological engineering
bioengineering
architectural engineering
biomedical engineering
medical engineering
ceramic engineering
chemical engineering
biomolecular engineering
chemical & biomedical engineering
civil engineering
geotechnical engineering
structural engineering
transportation engineering
highway engineering
water resource engineering
computer engineering
electrical engineering
communication engineering
electronics engineering
laser engineering
engineering mechanics
engineering physics
engineering sciences
environmental engineering
environmental health engineering

telecommunications
engineering
materials engineering
mechanical engineering
metallurgical engineering
mining &mineral engineering
naval & marine engineering
nuclear engineering
ocean engineering
petroleum engineering
systems engineering
textile engineering
polymer engineering
plastic engineering
construction engineering
forest engineering
industrial engineering
manufacturing engineering
operations research
surveying engineering
geological engineering
geophysical engineering
paper engineering
electromechanical engineering
mechatronics
robotics
automation engineering
biochemical engineering
engineering chemistry
biosystems engineering