Proceedings of the 2022 Pre-ICIS SIGDSA Symposium

Special Interest Group on Decision Support and Analytics (SIGDSA)

12-12-2022

Comparison of Federal and Private Sector Job Postings: A Data Science Term Frequency Analysis

Alexis M. Votto

Dylan Manuel

Rohit Valecha

H. Raghav Rao

Comparison of Federal and Private Sector Job Postings: A Data Science Term Frequency Analysis

Extended Abstract

Alexis M. Votto
University of Texas at San Antonio[1]
Alexis.Votto@utsa.edu

Dylan Manuel
University of Texas at San Antonio[2]
Dylan.Manuel@my.utsa.edu

Rohit Valecha, Ph.D.
University of Texas at San Antonio[1]
Rohit.Valecha@utsa.edu

H.Raghav Rao, Ph.D.
University of Texas at San Antonio[1]
HR.Rao@utsa.edu

1 Department of Information Systems and Cyber Security, Alvarez College of Business

2 Department of Computer Science, College of Science

Abstract

As Artificial Intelligence (AI) and Machine Learning (ML) capabilities expand across different industries, Human Resource (HR) professionals face the challenge of acquiring top talent to fill necessary vacancies within their organization. In data science, these vacancies differ across the private and federal sectors. Understanding which sector is more effective in communicating talent requirements has become imperative to maintain a competitive advantage in talent management. Thus, this research takes a novel approach to analyze job postings from the United States (US) federal government and US-based data science job postings within the private sector. We leverage a natural language processing model and term frequency-inverse document frequency (TF-IDF) analysis to analyze the job postings within our datasets. We then compare the results of our TF-IDF analysis to identify the representation of data science competencies within job postings and how the federal and private sectors differ.

Keywords

Human Resource Management, Term Frequency, Job Postings, Natural Language Processing

Introduction

With the evolution of Artificial Intelligence (AI) and Machine Learning (ML) saturating the workforce, companies face the challenge of hiring and internally securing the right talent to manage, develop, and operate these assets ethically. The United States Bureau of Labor Statistics projects a 36% growth of the data science career field between 2021 and 2031, which is much faster than the average for all other occupations within the federal and private sector. This evolution has consequently affected how human resource professionals manage the necessary skill demands that come with the evolution of technology and work environments. Historically, the federal sector relies heavily on contracted data science personnel rather than full-time staff, which has stemmed the Office of Personnel Management (OPM) to derive a new data science job series to better recruit more fulltime data science personnel. This shift in interest consequently creates a competitive environment for recruiting top-tier talent within the data science field. Thus, we seek to fortify recent research initiatives by diverting from the trend of analyzing specific skills desired of data scientists and investigate how the federal and private sector communicate their respective desires via natural language processing techniques to establish nomenclature and lexicons related to data science competencies derived from theory and industry practice. In doing so, this endeavor seeks to educate the industrial and academic community on how the private and federal sector thematically communicate their needs from data scientists rather than just analyze desire skills. By drawing attention to thematic differences within job descriptions, we seek to bolster recruitment strategies such that HRM professionals can remain agile when faced with competition.

Theoretical Background and Framework

With the recent explosion of academic and industry interest in data science and talent management, HRM literature has yielded a variety of analyses that have built a foundation for our research (Votto et al., 2021). Highlighting these impacts, we seek to convey how the contribution of our research is unique and differs from existing initiatives. For instance, Sibarani et al.'s (2017) efforts identify how AI/ML can be used to data science skills within job descriptions. Similarly, Heidarysafa et al. (2021) leverage these tools to investigate differences between data scientists, analysts, and machine learning engineers. Additionally, De Mauro et al. (2021) provide deeper insight into developing a standardized nomenclature for data scientist job responsibilities. There have also been research efforts which have identified a need for soft skills within this field (Gardiner et al., 2017; Halwani et al., 2022). We seek to build upon these authors by expanding upon the work of De Mauro et al. (2021) and Halwani et al. (2022) by exploring AI within HR responsibilities such as job descriptions of data scientists (expanding to analysists and machine learning specialists) to establish lexicons based on competencies derived from practice and theory. Specially, we extract five data science competencies from Deloitte's "Trustworthy AI™" multidimensional AI framework and IBM's Data Science Skills Competency Model. The five competencies are machine learning, ethics, artificial intelligence, statistics, and problem solving (Ammanath, 2022; Gottipati et al., 2021; IBM, 2020; Mökander and Floridi, 2021).

Methodology

We analyze the content of archival job postings from both the United States federal government (USAjobs) and the private sector (Glassdoor). We specifically investigated the primary job responsibilities and job description of 3,319 private sector job postings and 1,161 federal postings. Our methodology consisted of a 4-phased approach modeled after Meyer (2019). The first phase is the preparation phase which encompasses data collection methodologies and data pre-processing for the Word2Vec model. The second phase of this approach consists of organizing the data. Within this phase, we leverage the results from the preparation phase to build term-similarity dictionaries to guide our analysis. The third phase is the analysis phase which encompasses the conduction of a term-frequency inverse document frequency (TF-IDF) analysis. Within this step, we analyze the job postings' TF-IDF score. We incorporated the fourth phase, given the tools and steps to analyze the information. The last phase of this approach is where we convey the results of our analysis to address our research propositions. Phase-4 is otherwise known as the results section of this paper.

Results

Having identified that each dictionary was statistically significant for private and federal sectors, we conducted a TF-IDF analysis for each job posting relative to each data science competency dictionary for the federal and private sectors to measure comparisons. We identified the federal sector places a stronger emphasis action and duty centric verbiage when communicating data science requirements. Contrarily, the private sector is thematically more model and character centric within their lexicon. Our results indicate the private sector ranks problem-solving as the highest competency in the TF-IDF analysis, leveraging language related to character traits (honest rapport, enthusiasm, candor) whereas the federal sector ranks ethics as its strongest competency. The federal sector's thematically duty-bound lexicon for ethics leveraged words which indicate a responsibility of the employee (align, necessary, impact, prevent). To validate these differences, we conducted an ANOVA analysis to investigate the differences between each competency (Kim et al., 2005) and post-hoc T-Test to investigate the mean differences between the two sections (private and federal sector). According to the TF-IDF ANOVA results there is sufficient evidence that supports the mean TF-IDF differs among the five competencies for federal and private job sectors at a .05 level of significance. Through our post-hoc TF-IDF means comparison, we identified and validated the ethics dictionary has the highest TF-IDF difference between the federal and private sector. This indicates the ethics dictionary was the least represented in one (private sector) and most represented in the other (federal sector). Contrarily, the problem-solving dictionary had the smallest difference. This implies that both the federal and private sector placed a similar emphasis on utilizing vocabulary within their respective problem-solving dictionaries.

Discussion

By shining a light on these differences, employers can leverage the identified dictionaries to better communicate job requirements based on recent job posting trends for data scientists. This insight can potentially better posture applicants for their resume and application development, as well as employers in job description development. Our analysis encourages future research endeavors to investigate the relationship between the ethical expectations of employees and the semantics within job descriptions. Future research could explore how to leverage the established dictionaries to build more robust job descriptions and measure how successful it is in attracting applicants or existing employees. Considering these future research themes, we acknowledge that it is essential to identify the limitations of this research. Although systematic, the methodology for data collection limited us in the variables we could explore. Had we had access to temporal considerations (how long the posting was open) or attractiveness of the posting (number of applications received), we could have dived deeper into how the dictionaries could affect the recruitment efforts of the federal and private sectors.

Conclusion

Within this analysis, we took a novel approach to analyze the content within job postings to understand better how the US federal government may differ from private companies soliciting on GlassDoor. Where previous literature has homed explicitly in on skill requirements, we wanted to divert from this trend to focus on the content within the job descriptions to understand better how the federal government and private sector communicate to prospective data scientists. By leveraging IBM Data Science competencies (IBM, 2020) and Deloitte's "Trustworthy AI™" framework (Ammanath, 2022), we leveraged established data science competencies to build dictionaries containing words that were most like them. Through this extensive analysis, we identified that the federal sector's most vital dictionary was ethics, and the private sector favored problem-solving. We hope the proposed data competency model and established dictionaries can help HR professionals build more robust job descriptions for data scientists.

References

Ammanath, B. 2022. *Trustworthy Ai: A Business Guide for Navigating Trust and Ethics in Ai*. John Wiley & Sons.

Deloitte. 2020. "Trustworthy Artificial Intelligence (Ai)™," *Deloitte United States*).

De Mauro, A., Greco, M., Grimaldi, M., and Ritala, P. 2018. "Human Resources for Big Data Professions: A Systematic Classification of Job Roles and Required Skill Sets," Information Processing & Management (54:5), pp. 807-817.

Gardiner, A., Aasheim, C., Rutner, P., and Williams, S. 2018. "Skill Requirements in Big Data: A Content Analysis of Job Advertisements," *Journal of Computer Information Systems* (58:4), pp. 374-384.

Gottipati, S., Shim, K. J., and Sahoo, S. 2021. "Glassdoor Job Description Analytics – Analyzing Data Science Professional Roles and Skills," *2021 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1329-1336.

Halwani, M. A., Amirkiaee, S. Y., Evangelopoulos, N., and Prybutok, V. 2021. "Job Qualifications Study for Data Science and Big Data Professions," Information Technology & People (35:2), pp. 510-525.

Heidarysafa, M., Kowsari, K., Bashiri, M., and Brown, D. E. 2021. "Toward a Knowledge Discovery Framework for Data Science Job Market in the United States." arXiv.

Ibm. 2020. "The Data Science Skills Competency Model," *IBM Corporation*), p. 12.

Meyer, M. A. 2019. "Healthcare Data Scientist Qualifications, Skills, and Job Focus: A Content Analysis of

Mökander, J., and Floridi, L. 2021. "Ethics-Based Auditing to Develop Trustworthy Ai," *Minds and*

Opm. "Position Classification Flysheet for Data Science Series, 1560,"), p. 19.

Sibarani, E. M., Scerri, S., Morales, C., Auer, S., and Collarana, D. 2017. "Ontology-Guided Job Market Demand Analysis: A Cross-Sectional Study for the Data Science Field," *Semantics2017: Semantics 2017 - 13th International Conference on Semantic Systems*: ACM, pp. 25-32.

Votto, A. M., Valecha, R., Najafirad, P., and Rao, H. R. 2021. "Artificial Intelligence in Tactical Human Resource Management: A Systematic Literature Review," International Journal of Information Management Data Insights (1:2), p. 100047.