EPiC Series in Built Environment

Volume 3, 2022, Pages 166–173

ASC2022. 58th Annual Associated Schools of Construction International Conference

Evolution of Patterns in Inclusion, Diversity, Equity, and Accountability (IDEA) and Race, Ethnicity, and Language (REAL) in Construction Job Market using Web Scraping and Text Mining Techniques

Heung Jin Oh and Baabak Ashuri, Ph.D.
Georgia Institute of Technology
Atlanta, Georgia

Mohsen Shahandashti, Ph.D., P.E.
University of Texas at Arlington
Arlington, Texas

The primary objective of this research is to characterize job market trends and identify patterns in Inclusion, Diversity, Equity, and Accountability (IDEA) and Race, Ethnicity, and Language (REAL) in the construction industry. Workforce IDEA and REAL are major concerns in the construction industry. The construction industry has challenges of embracing the workforce IDEA and REAL such as age, language, sexual orientation, and disability. IDEA and REAL have impacts on hiring as well. However, there is a lack of data and research related to IDEA and REAL for the construction industry. In this respect, this research collects construction job market data using web scraping. Text mining techniques are then applied to detect words related to IDEA, REAL, job types, and wage changes during the web scraping period. We devise a new data-driven text mining model including the web scraping technique to provide market trends and identify patterns related to IDEA and REAL in the construction industry. This research will contribute to the body of knowledge of capturing the data regarding IDEA and REAL by harvesting large-scale data across disciplines in a searchable and organized structure.

Key Words: Construction workforce, Job market, Diversity, Inclusion, Web scraping, Text Mining

Introduction

Inclusion, Diversity, Equity, and Accountability (IDEA) play a significant role in any workforce and are important indicators for the sustainable development of an industry (Karakhan, Gambatese, Simmons, & Al-Bayati, 2021). IDEA-related policies or systems for a specific group can cause changes in the productivity of the overall workforce (Toohey, 2009). The construction industry is a labor-intensive industry; for example, contractors such as mechanical or electrical trades account for labor costs up to 50% of the total cost of projects (Hanna et al. 2005). Although the importance of the labor force in the construction industry cannot be overemphasized, the construction industry is still far from achieving IDEA (Powell & Sang, 2013) which can lead to changes in labor productivity. Previous projects and researches have been focused on organizational sustainability or worker safety and health rather than IDEA issues (Gambatese, Karakhan, & Simmons, 2019). Race, Ethnicity, and

Language (REAL) also has a significant impact on the workforce in hiring, such as determinants of employers' hiring decisions and hiring discrimination (Baert, 2018).

The structure of the workforce in the construction industry is quite unique compared to other industry. The proportion of women in the construction industry is 10.9%, which is significantly lower than the average (46.8%), and the proportion of Hispanic or Latino is 30.0%, which is higher than the average (17.6%) in the overall industry (U.S. Bureau of Labor Statistics, 2020). Some studies conducted interviews and surveys for male-dominated structure in construction. For example, Galea et al. (2015) had interviews with senior managers and analyzed formal policies. The study concluded that previous initiatives and policies are more likely to promote increasing the numbers of women in construction rather than gender practices and outcomes (Galea, Powell, Loosemore, & Chappell, 2015). On the other hand, some studies are focused on low-literacy and low-English-proficiency workers, including Hispanic workers. A case study suggests scenario-based 3D training materials to reduce construction fall fatalities for the workers who are not proficient in English (Lin, Lee, Azari, & Migliaccio, 2018). However, previous studies have limited to particular problems rather than overall impacts of IDEA or REAL in the job market and small size of data collected by interviews, surveys, or case studies. There is also a lack of research directly on IDEA or REAL related to hiring.

In this respect, this research aims to investigate impacts of IDEA and REAL on the construction job market by analyzing job description data. Today, online job postings contain various contents such as job requirements, company descriptions, or what kind of company values they pursue. This research assumed that online job postings contain related words to achieve IDEA and REAL in the construction industry. Publicly available construction job information posted on Indeed.com is utilized to collect throughout the United States. Texts in job descriptions are analyzed by using text mining techniques. Therefore, this study identifies the trends towards how much IDEA and REAL are considered in the construction job market. The findings will contribute to understanding and developing IDEA and REAL in the construction industry.

Methodology

This research has the following steps in Figure 1. Data collection is mainly done by web scraping using bots. The code for bots is formulated based on R programming language. The data contains job postings from 04/25/2020 to 05/07/2021. Some missing periods due to changes on the website, problems with scraping, or tests at the beginning level are not included in the research. The number of raw data points is about 13.5 million. This study also includes unemployment rates, job types in the construction industry from BLS (U.S. Bureau of Labor Statistics, 2021), region and division data from the U.S. Census (U.S. Census Bureau, 2010).

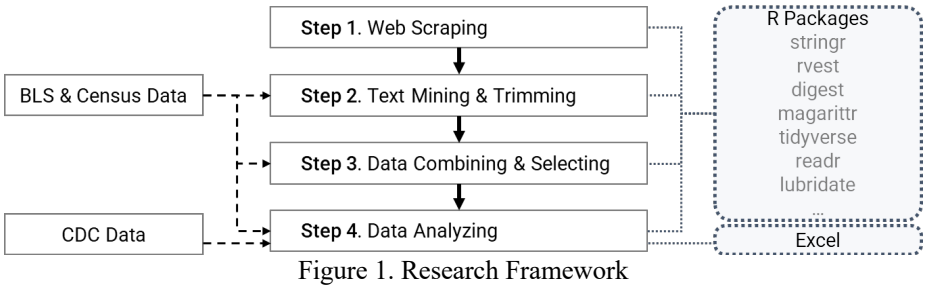

Figure 1. Research Framework

Data processing is also formulated using R programming language. Raw data is summarized in Table 1. To process the raw data from web scraping, this research first defines four forms of wages in three

terms, Min, Max, and Fixed wages in Table 2. Based on the criteria from BLS data, this study calculates and unifies units of wages as 40 hours in a week and 52 weeks in a year. Job postings are classified based on keywords of IDEA and REAL. Job title, job summary, and job description are also used for classifying job postings when matching job types. Each data point is also categorized by four regions and nine divisions (U.S. Census Bureau, 2010). Google API is used to match location information from job postings and the regions and divisions. In addition, to refine the data, this research combines overlapped data points from different dates, deletes data not including critical information such as wages, and excludes data not classified by job types and outliers. 13,450,369 data points are converted into 371,789 data points on these processes. Lastly, the data is converted to .csv file format for data analysis.

Table 1

Raw data summary

Data	Source	Description
1. Web Scraping	Indeed	Date, state, location(mostly city level), salary, number of days from upload, job title, summary, description
2. Job Types	BLS	Based on NAICS
3. Wage by Job Types	BLS	Most of job types in the construction industry
4. Region and Divisions	Census	Four regions and nine divisions in the United States
5. Unemployment Rate	BLS	Monthly rate

Table 2

Converting wages into Min/Max/Fixed wage

Examples	Min Wage	Fixed Wage	Max Wage
$10 ~ $20 / hour	$10		$20
From $10 / hour	$10		
Up to $20 / hour			$20
$15 / hour		$15	

Classification keywords for IDEA and REAL are as follows:

- IDEA: diversity, equity, inclusion, creed, religion, gender; and
- REAL: spanish, chinese, mandarin, tagalog, filipino, vietnamese, french, cajun, italian, german, polish, greek, portuguese, japanese, yiddish, korean, russian, caucasian, hispanic.

Classification keywords for Job Types are as follows:

- labor, helper | assistant, painter, operator, drywall | ceiling | taper | hanger, glazier, insulat, inspector | examiner, electric, hazardous material | asbestos | lead, mason | stone, carpent, floor | tile | marble | carpet, iron | steel | rebar, sheet metal | sheetmetal, plumb | steamfit | fitter, solar | pv, roof, elevator | escalator, boiler

After processing the data in Table 1, the final data set has 50 columns (Table 3). R programming language and Excel are used to analyze the final data set.

Table 3

Final data summary

Column Name	Format and Unit	Data Source (Table 1)
UniqueID	Char.	Data.1: uniqueid (overlapped combined)
Startdate	Date	Data.1: scraped date, number of days from upload
State	Char., 2digits	Data.1: state
Year-month	Char., 6digits	Data.1: combine year and month
Year-week	Char., 6digits	Data.1: combine year and week
IDEA words	0, 1, or NA	Data.1: job description
REAL words	0, 1, or NA	Data.1: job description
Job types (27 Columns)	0, 1, or NA	Data.1: detect from job title, summary, description Data.2: job types
Hourly mean Hourly max Hourly fixed	$/hour	Data.1: extract from salary, trimming Data.3: 40 hours in a week, 52 weeks in a year
Region and Divisions (13 Columns)	0, 1, or NA	Data.1: state Data.4: regions and divisions

Results

This study finally analyzes job postings related to IDEA and REAL. It first analyzes and visualizes the trends of texts in job postings. It also classifies them into IDEA and REAL and analyzes the trend of wage included in the job postings accordingly. Also, the job postings are classified by job types and regional information. These results will confirm the impact of IDEA and REAL on the overall construction job market by comparing average wages from job postings.

To find which words are mainly included in IDEA and REAL (Figure 2), R packages such as tm, SnowballC, wordcloud, RColorBrewer are used for the analysis and visualization. Collateral texts including numbers, punctuations, extra spaces are carried out in this process. Words among top 50 for IDEA and REAL are as follows:

- IDEA: disability, electrical, requirements, status, employee, federal, color, race, religion, perform, opportunity, veteran, information, equal, origin, national; and
- REAL: commercial, customer, English, experienced, fulltime, good, high, school, hour, looking, management, office, quality, remotely, responsible, Spanish; and
- In Both IDEA and REAL: ability, applicants, company, construction, equipment, experience, job, must, required, safety, team, time, will, work, years, etc.

Figure 2. Visualization for job postings with IDEA (Left) and REAL (Right)

Data extracted from job postings also provides average wages with IDEA and REAL. The monthly averages of max and min wages show similar trends. Both wages have decreased from April 2020 to June 2020, and they recovered during the next five months. In case of comparing all average wages with IDEA and REAL data, the average wage with IDEA shows higher than the others (Figure 3).

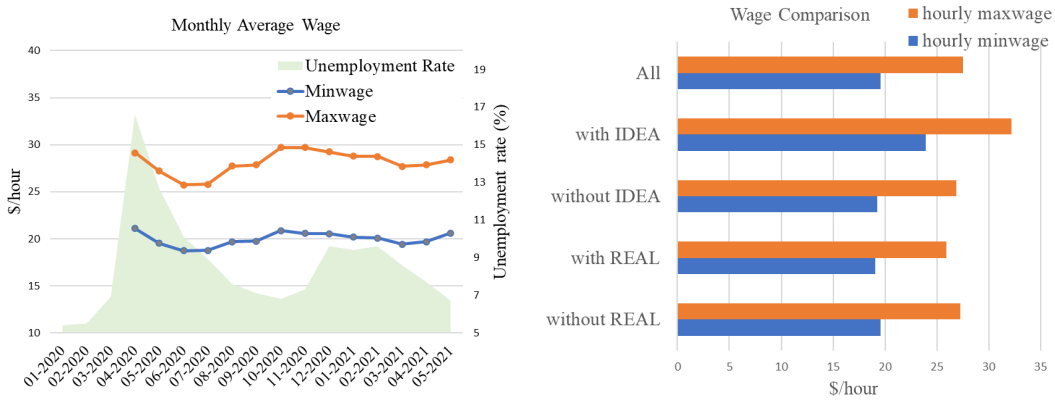

Figure 3. Monthly average wages and wage comparisons with/without IDEA and REAL

The average wage with IDEA shows higher numbers than without IDEA in every month for both max and min wage. On the other hand, monthly average wage with REAL decreases faster than monthly average wages without REAL from April 2020 to June 2020 and recovers faster during the next months (Figure 4).

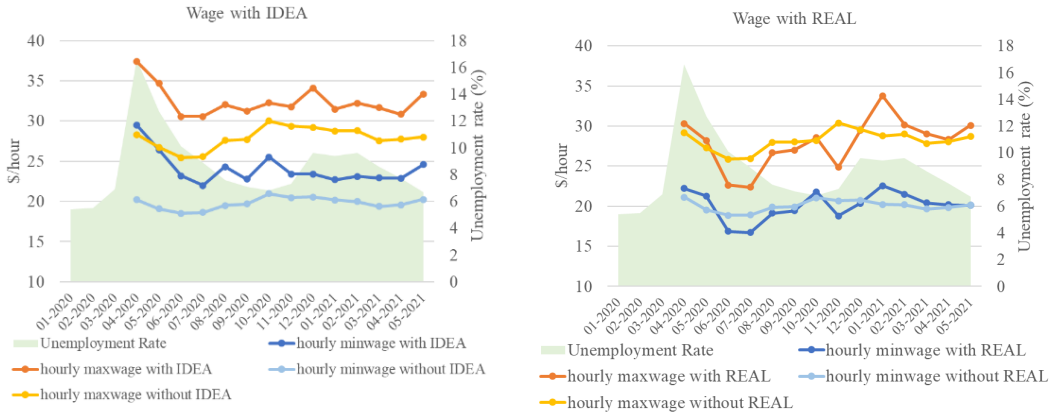

Figure 4. Monthly wage changes by IDEA (Left) and REAL (Right)

The averages of min and max wages from job postings by job types are compared in Figure 5. The average wages with IDEA show higher wages in most of job types except for operator, glazier, and flooring compared to average wages. However, the average wages with REAL show lower wages in many job types, such as helper, operator, insulator, electrician, and hazardous. Only painter, ceiling, mason, ironsteel, PV (photovoltaic), and boiler show higher wages with both REAL and IDEA.

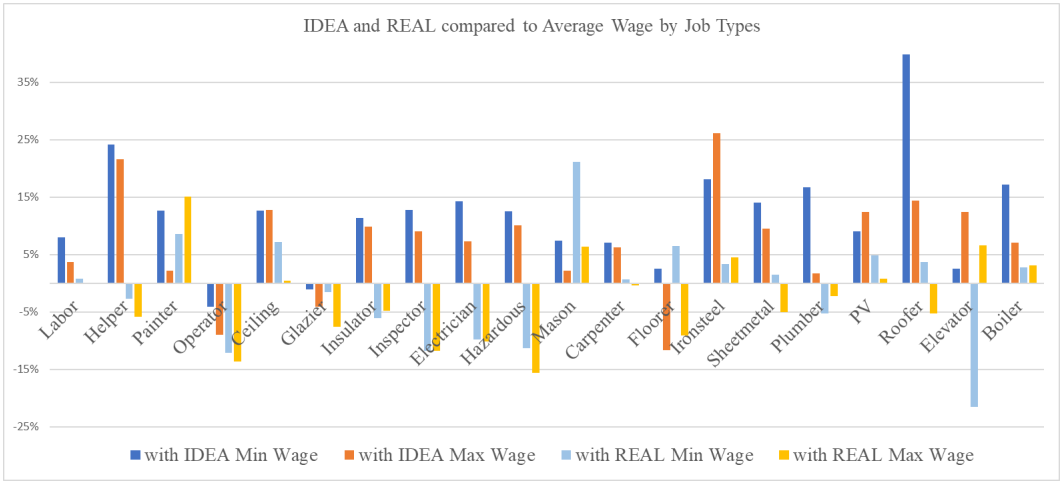

Figure 5. with IDEA and REAL wages compared to average wage by job types

By four regions and nine divisions, job postings with IDEA offer higher wages than average wages in all the regions and divisions. In case of job postings with REAL, wages are only higher in the Midwest region and the West North Central and East South Central divisions (Figure 6).

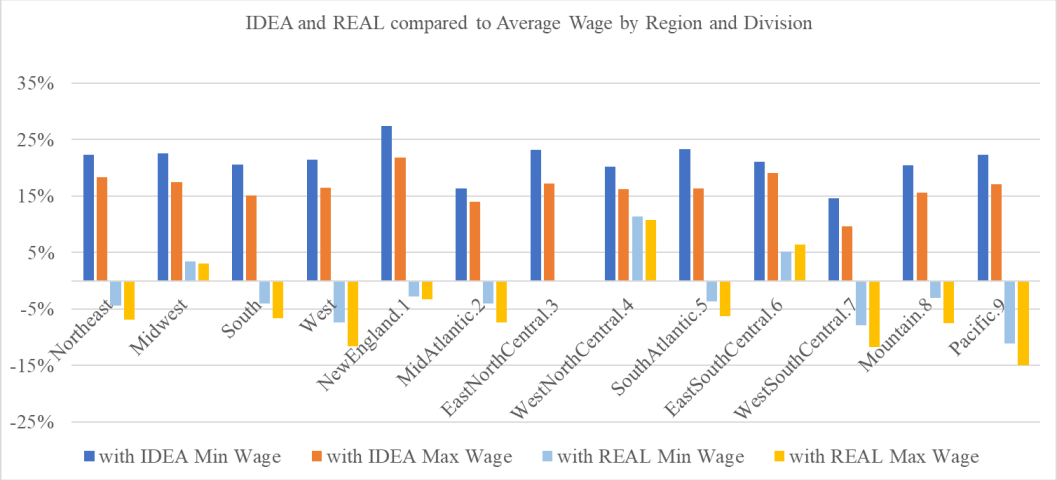

Figure 6. with IDEA and REAL wages compared to average wage by region and division

Conclusions

Web scraping and text mining techniques have been conducted to identify patterns in Inclusion, Diversity, Equity, Accountability (IDEA), Race, Ethnicity, and Language (REAL) for the construction job market in this research. This study also finds words that frequently appear in job postings which include words related to IDEA and REAL. In addition, the study analyzes trends including wage, monthly data, job types, and regions and divisions. Overall, job postings with IDEA words offer higher wages compared to average wages, but job postings with REAL do not (Figure 3). When the unemployment rate increases, average wages with REAL decrease more rapidly than wages with IDEA (Figure 4). Also, the impact of IDEA and REAL is different by job types and region and divisions (Figure 5 and 6). Especially in some job types and region and divisions, job postings with REAL offer lower wages compared to average wages. Job postings with REAL offer relatively unfavorable wages in many cases in this study. It is encouraging that job postings with IDEA offer higher wages, but we should pay more attention to REAL.

Findings in this paper will support the job market, benefitting both employers and employees and contribute to the body of knowledge for the sustainable workforce developments in the construction industry. The study detects where the construction industry should take into account IDEA and REAL to achieve balanced development including job types and regions and divisions. In addition, the data-driven approach using web scraping and text mining assists in the continuous monitoring of trends in workforce IDEA and REAL in the construction industry. Although this study has limitations in terms of whether data from web scraping can represent the entire construction job market, it has advantages in using far more data about workforce IDEA and REAL than other studies. This research also has limitations in that it did not consider the impacts of the pandemic on the construction job market. We plan to consider the impacts of COVID-19 with the data provided by the Centers for Disease Control and Prevention (CDC, 2021) for future work. This data-driven approach in this study will also enable us to integrate construction labor market data with other big data and contribute to sustainable workforce development in the construction industry.

Acknowledgement

This paper is based upon work supported by the National Science Foundation under Grant No. 2035198 and 2035299.

References

Baert, S. (2018). Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (pp. 63–77). Springer, Cham. https://doi.org/10.1007/978-3-319-71153-9_3

CDC. (2021). CDC COVID Data Tracker. Retrieved June 4, 2021, from Centers for Disease Control and Prevention website: https://covid.cdc.gov/covid-data-tracker/#datatracker-home

Construction and Extraction Occupations : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics. (2021). Retrieved June 4, 2021, from U.S. Bureau of Labor Statistics website: https://www.bls.gov/ooh/construction-and-extraction/home.htm

Galea, N., Powell, A., Loosemore, M., & Chappell, L. (2015). *Construction Management and Economics Designing robust and revisable policies for gender equality: lessons from the Australian construction industry*. https://doi.org/10.1080/01446193.2015.1042887

Gambatese, J. A., Karakhan, A. A., & Simmons, D. R. (2019). *Development of a Workforce Sustainability Model for Construction*.

Karakhan, A. A., Gambatese, J. A., Simmons, D. R., & Al-Bayati, A. J. (2021). Identifying Pertinent Indicators for Assessing and Fostering Diversity, Equity, and Inclusion of the Construction Workforce. *J. Manage. Eng. J. Manage. Eng*, 37. https://doi.org/10.1061/(ASCE)

Lin, K.-Y., Lee, W., Azari, R., & Migliaccio, G. C. (2018). Training of Low-Literacy and Low-English-Proficiency Hispanic Workers on Construction Fall Fatality. *Journal of Management in Engineering*, *34*(2), 05017009. https://doi.org/10.1061/(asce)me.1943-5479.0000573

Powell, A., & Sang, K. J. C. (2013). *Construction Management and Economics Equality, diversity and inclusion in the construction industry*. https://doi.org/10.1080/01446193.2013.837263

Toohey, T. (2009). *Australia's hidden resource: the economic case for increasing female participation*. Goldman Sachs JBWere.

U.S. Bureau of Labor Statistics. (n.d.). Unemployment Rate - Construction Industry, Private Wage and Salary Workers. Retrieved June 1, 2021, from U.S. Bureau of Labor Statistics website: https://data.bls.gov/timeseries/LNU04032231?amp%253bdata_tool=XGtable&output_view=data&include_graphs=true

U.S. Bureau of Labor Statistics. (2020). Labor Force Statistics from the Current Population Survey. *Labor Force Statistics from the Current Population Survey*. Retrieved from https://www.bls.gov/cps/cpsaat18.htm

U.S. Census Bureau. (2010). Census regions and divisions of the United States. Retrieved June 1, 2021, from U.S. Census Bureau website: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf