Measuring Corporate Human Capital Disclosures: Lexicon, Data, Code, and Research Opportunities*

Elizabeth Demers§
Victor Xiaoqi Wang€
Kean Wu£

December 2023

Abstract

Human capital (HC) is increasingly important to corporate value creation. Unlike other assets, however, HC is not currently subject to well-defined measurement or disclosure rules. We use a semi-supervised machine learning algorithm (word2vec) trained on a confirmed set of HC disclosures to develop a comprehensive list of HC-related keywords classified into five subcategories (diversity, equity, and inclusion; health and safety; labor relations and culture; compensation and benefits; and demographics and other) that capture the multidimensional nature of HC management. We share our lexicon, machine-readable corporate HC disclosures, and the Python code used to develop the lexicon, and we provide a detailed example of the code’s application. Researchers can use our HC lexicon (or modify the code to capture another construct of interest) with their own samples of corporate communications to address HC questions of interest. We close with a discussion of future research opportunities related to HC management and disclosure.

Keywords: Human capital, corporate disclosure, lexicon, word2vec, BERT, Python code, textual analysis

* The authors have no conflicts of interest to declare. All data is from public sources.
§ School of Accounting and Finance, University of Waterloo, Ontario, Canada. elizabeth.demers@uwaterloo.ca.
€ College of Business, California State University, Long Beach, CA, USA. victor.wang@csulb.edu.
£ Saunders College of Business, Rochester Institute of Technology, Rochester, NY, USA. kwu@saunders.rit.edu.
Measuring Corporate Human Capital Disclosures: Lexicon, Data, Code, and Research Opportunities

I. INTRODUCTION

In the modern, knowledge-based economy, human capital is often considered to be a company’s most important asset (Batish et al. 2021) and a key driver of firm value (e.g., Edmans 2011), even while sustainable investing and investor awareness of social justice issues are also on the rise. Investor demand for information about how firms manage their human capital (HC) has therefore become correspondingly important.¹ But measuring textual HC disclosures is not an easy task. Human capital management (HCM) is multi-faceted, including aspects such as recruitment, talent development, health and safety, compensation management, and fostering an inclusive environment. Prior studies related to HC disclosures rely on short lists of keywords or manual coding schemes that are often not adequate for capturing the multi-dimensionality of HCM practices, nor are they suitable for large-sample analyses. In this paper, we develop a comprehensive list of HC-related keywords by using a semi-supervised machine learning algorithm, the word2vec model (Mikolov et al. 2013). Researchers can use our lexicon to identify and measure HC disclosures with a view to examining many timely research questions related to corporate HC management.

The first and most important step involved in quantifying textual HC disclosures is to reliably identify HC-related contents. Before the SEC mandated the human capital (HC) disclosures that are now provided in a separate section within Item 1 of Form 10-Ks, firms

¹ For example, in their survey of institutional investors, Morrow Sodali (2019) found that 83% of respondents indicated that the ESG topic most in need of improved disclosures was human capital. See also discussions about the petition from the Working Group on Human Capital Accounting Disclosure to the SEC (Posner 2022).
voluntarily provided HC-related disclosures in various documents or venues such as regulatory filings, sustainability reports, earnings conference calls, and company websites. Moreover, such disclosures were often dispersed and embedded in other non-HC-specific content. We capitalize on the SEC’s new disclosure requirements to train our model on these now reliably identifiable HC-specific disclosures. To begin, we hand collect HC disclosures from 10-Ks filed by approximately 4,000 firms during the first year of the new regulation, from November 2020 to November 2021. Training our model on this corpus of unambiguously HC-related documents consisting of almost two million words ensures that our lexicon has high construct validity. Our lexicon covers five broad topics in HCM, including “Diversity, Equity, and Inclusion (DEI)” (253 terms), “Health and Safety” (227 terms), “Labor Relations and Culture” (362 terms), “Compensation and Benefits (283 terms),” and “Demographics and Other” (160 terms). For “Health and Safety,” we further classify the keywords into those that are specific to COVID-19 (70 terms) and those that are more generic (157 terms).

With a view to contributing to recent initiatives that promote the use of novel datasets to answer interesting research questions (e.g., Hoitash and Hoitash 2022; Wang and Wang 2022; Abu-Khadra and Olsen 2023), we share our lexicon, together with the approximately 4,000 machine-readable corporate textual disclosures and Python code used to develop the keyword list. We also provide sample code that can be used together with the word list to extract HC measures from any other user-provided collection of textual documents. Researchers can use our lexicon and/or adapt our codes to create alternative HC measures that suit their particular research setting. Using our code, researchers can also apply the word2vec algorithm to develop word lists for other (i.e., non-HC) constructs of interest by supplying their own list of initial seed words to the algorithm. The word2vec algorithm is a promising new technique in textual analysis as it
overcomes a major weakness of the traditional dictionary-based approach, which often relies on a non-comprehensive list of keywords identified by the researcher. The latter may be based upon the researcher’s own limited perceptions, training, and experiences, or the selection of source documents (e.g., textbooks), and therefore may not capture the range of terminology in current use by a varied set of corporate reporters. Our approach overcomes these limitations.

Recent developments in machine learning provide alternative ways to identify and classify text, such as using transformer-based Large Language Models (LLMs) like BERT (Devlin et al. 2019). However, the application of such models to a specific domain often necessitates pre-training a new model on a large corpus of domain-specific text and/or fine-tuning the pre-trained base model using labeled datasets. Two notable examples are FinBERT (A. H. Huang, Wang, and Yang 2023), which is pre-trained on finance text, and ClimateBERT (Webersinke et al. 2022), which is pre-trained on climate disclosure text. The HC disclosures that we collect and share with this study can serve as valuable resources for pre-training a BERT model specific to HC disclosures or for fine-tuning the base BERT model for HC disclosure classification. In addition, our keyword list can also serve as a useful tool for facilitating pre-training or fine-tuning by identifying HC-specific documents or pinpointing human-labeled samples that may require additional scrutiny. To illustrate the potential use of our data for training or fine-tuning a domain-specific LLM, we fine-tune the base BERT model for identification of HC content using the confirmed HC disclosures that we collected as labeled datasets.

In the remainder of the paper, we begin by providing a review of the prior HC literature, with a special focus on the measurement of HCM practices and HC disclosures. This review reveals that the literature lacks a comprehensive measure for quantifying textual HC disclosures or otherwise gauging a firm’s HCM practices for a large cross-section of firms. We then elaborate
on how we develop our lexicon as a new alternative for measuring corporate HC disclosures and HCM practices, and demonstrate how to use the lexicon by constructing HC disclosure measures for a large collection of proxy statements containing more than 3.3 billion words. We further illustrate an alternative way of using our data to classify HC disclosures by fine-tuning the base BERT model, and discuss the advantages and disadvantages of each of keyword-based and machine-learning-based classifications. Finally, we propose a number of avenues for future HC research that could be pursued using our lexicon and our data.

II. LITERATURE REVIEW

Human capital is a topic of interest to researchers in many disciplines, including management, strategy, and organizational behavior, amongst others. Within the realm of accounting and finance, two streams of literature are noteworthy: i) the implications of HCM practices for the value creation process; and ii) corporate disclosure decisions related to their HCM. These two streams of literature are interconnected since firm disclosures offer an important channel for outsiders to understand a firm’s HCM practices and their implications for firm value.

In the context of value creation, prior studies find that various aspects of HCM practices contribute to firm value, including compensation (Rayton 2003), employee satisfaction (Edmans 2011; Green et al. 2019), and training (Molina and Ortega 2003). Because corporate disclosures of HCM practices are scant and are often in a format that makes them hard to collect and process (in the rare instance where they were available prior to the new regulation), most previous studies rely on external sources of information to gauge a firm’s HCM practices, including “best employer” ratings (e.g., Edmans (2011)) or crowd-sourced employee ratings from Glassdoor (e.g., Green et al. (2019)). Other researchers overcome the challenge imposed by lack of tabulated
disclosures by using survey-based measures, although such studies are necessarily limited to the firms that respond to the survey, often resulting in response bias and small sample data analysis constraints.²

The second stream of HC literature examines corporate voluntary disclosure practices. These studies largely use disclosure indexes that are manually constructed by researchers who read the disclosure text and assign a value of one or zero to indicate the presence or absence of a particular disclosure item. This traditional method of manual content analysis suffers from several weaknesses, including that the disclosure items of which the indices are composed are subjectively chosen, the manual coding of these items is time-consuming, and as a result, the methodology is typically only useful for small sample studies and produces results that are often hard to replicate.

For example, Vithana, Soobaroyen, and Ntim (2021) create disclosure indexes for UK FTSE 100 firms according to a list of 22 disclosure items that include “Employee health and safety,” “Employee diversity and equity,” and “Employee involvement and engagement,” among others. These authors find that HC disclosures are inadequate in both breadth and depth for the largest FTSE firms, but that firms embracing an employee relation ideology of unitarism tend to provide a higher level of disclosure. Another study by McCracken et al. (2018) uses a sample of FTSE 100 firms to measure the level of HC disclosure using a list of items covering four broad categories, including “Knowledge, Skills, and Abilities (KSA),” “Human Resource Development (HRD),” “Employee welfare,” and “Organizational justice and equity.” Each of these four categories consists of a slightly expanded list of terms. For example, HRD includes “apprenticeships,” “career development,” “graduates,” “internships,” “succession planning,”

² Bernstein and Beeferman (2015) provide a summary of this literature.
“talent management and training,” and other similar terms. They find that the top 100 UK companies significantly increased their level of HC disclosure after the UK Companies Act 2006 was amended in 2013 to require companies to report on their human capital.

Using the Canadian setting, Cormier et al. (2009) examine HC disclosures on the corporate websites of 155 non-financial firms included in the Toronto Stock Exchange S&P/TSX Index. They find that voluntary HC disclosure reduces information asymmetry and that better corporate governance leads to a higher level of disclosure. They measure HC disclosures using a coding instrument that includes 33 items such as “hiring,” “training,” “performance assessment,” and “employee satisfaction.”

Some studies examine HC disclosures in the broader context of a firm’s sustainability reporting. For example, Ehnert et al. (2016) use manual coding to examine the extent of sustainable Human Resource Management (HRM) reporting for Forbes Top 250 global companies by using the framework of the Global Reporting Initiative (GRI). This framework covers quantitative and qualitative factors such as employee demographics, employee turnover, labor relations, occupational health and safety, training and education, and diversity and equal opportunity. Somewhat surprisingly, they find that these large firms provide disclosures on HRM performance that are commensurate with those related to environmental performance, and that international differences in HRM disclosures are not as large as expected.

A few studies examine corporate disclosures related to intellectual capital or property, which includes human capital. For example, Bontis (2003) examines the intellectual capital disclosures of Canadian companies and finds that the propensity for disclosure is surprisingly low during the sample period; only 74 instances of disclosure were found among 10,000 annual reports. He uses a list of 39 terms, about one third of which relate to human capital (e.g., “human capital,”
“human assets,” and “employee skill”). In a related very small sample study, Sujan and Abeysekera (2007) examine the intellectual property reporting practices of the 20 largest Australian firms by market capitalization. Using traditional manual content analysis with a coding scheme, they find that HC disclosures in annual reports were mostly qualitative, and that such disclosures lacked consistency across firms even though there was an observable trend towards more quantitative disclosures.

The SEC’s 2020 mandate on HC disclosures has generated tremendous interest in this topic, as numerous recent studies examine mandatory or voluntary HC disclosures in the U.S. setting. For example, Batish et al. (2021) manually read the HC disclosures provided by the 100 largest firms under the new regulation and subjectively conclude that these early disclosures are too generic to be useful to investors. Zhang (2022) examines the determinants and consequences of voluntary HC disclosures by using machine learning to extract HC disclosures from annual reports in a period before the new SEC regulation took effect. Demers, Wang, and Wu (2023) provide a comprehensive analysis of all available 10-K HC disclosures from the first two years of the new regulation. They find that the attributes of these disclosures (e.g., length, specificity, and tone) vary significantly across firms, and that the disclosures overall do not appear to be sufficiently informative due to their lack of specificity and numerical intensity. Larger firms are found to have better disclosures, whereas firms for which HC is more strategically important do not. Time trends suggest that there are learning effects (e.g., later filers appear to learn from earlier filers, and firms update their filings in the second year of the regulation), but also that there is reversion towards the mean as firms with weaker disclosures (e.g., shorter, less specific, less numerically intense, or less readable) improve their disclosures while previously better disclosers regress on these characteristics. Arif, Yoon, and Zhang (2022) find that the amount of HC
disclosure in 10-K filings is negatively associated with contemporaneous employee turnover, suggesting that such disclosures provide new information about a company’s human capital management practices. They also find that both equity markets and bond markets react to HC disclosures, although they do so differently, in ways that are probably explained by the respective investor groups’ different pay-off functions. Using manual content analysis on a sample of S&P 500 firms, Michaelides and Vafeas (2023) find that Chief Human Resource Officers (CHROs) play a significant role in determining the quality of HC disclosures, especially when the CHROs are more powerful. However, CHROs have a smaller impact when they come from underrepresented groups, such as when they are women or ethnic minorities.

A few recent studies focus on particular aspects of HC disclosure such as gender diversity (Liang et al. 2022), attraction and retention (Haslag, Sensoy, and White 2022), COVID-19 related issues (Mayew and Zhang 2022), and quantitative metrics (Bourveau et al. 2022). Liang et al. (2022) obtain their measures of gender diversity from external data providers (Bloomberg and Revelio Labs). They find that companies with a higher percentage of female employees are more likely to voluntarily disclose their employees’ gender diversity. Haslag, Sensoy, and White (2022) identify employee attraction and retention disclosures in 10-K filings by requiring a sentence to contain at least one word from each of two researcher-constructed word lists. They find that firms adjust their HC disclosures in response to changes in their workforce. Mayew and Zhang (2022) identify COVID-19 related HC disclosures by searching a few pandemic-related keywords in the firm’s HC disclosures. They find that the amount of such disclosures is positively related to employee satisfaction with firms’ response to the pandemic. Bourveau et al. (2022) manually collect quantitative metrics of HCM practices from annual reports in a sample period surrounding
the SEC’s 2020 regulation and find that there is a significant increase in quantitative disclosures in the post-regulation period.

In summary, the literature related to the role of HC in the value creation process largely uses HC measures obtained from external sources to gauge the impact of a firm’s HCM practices on firm performance. The literature related to corporate HC disclosure decisions, in turn, tends to use relatively short and non-comprehensive researcher-defined lists of select items or terms. These lists vary from study to study and this approach is not well-suited to large-sample analyses. Zhang (2022) is an exception to this in that she uses a machine learning approach that has the potential to create a list of comprehensive keywords for capturing various dimensions of a firm’s HCM practices. Her model is trained on a broad set of mostly non-HC corporate textual disclosures, however, which can significantly reduce the construct validity of the resulting lexicon. Our approach overcomes all these limitations by applying machine learning to a large sample of recent HC-specific corporate disclosures to develop an extensive HC-related dictionary of contemporary relevance with high construct validity.

In the next section, we discuss how we construct our list of HC-related keywords by applying a machine learning algorithm to nearly 4,000 corporate 10-K HC disclosures filed during the first year of the 2020 SEC regulation.

III. METHODOLOGY AND LEXICON

Word2Vec

We use word2vec, a word embedding model proposed by Mikolov et al. (2013), to construct a comprehensive list of HC keywords. Based on the assumption that words occurring in a similar context (i.e., surrounding words) tend to have similar meanings (Harris 1954), this
algorithm uses a neural network to learn word associations from a large collection of text. By representing each word as a numeric vector that captures the semantic qualities of that word, it allows the relationship between words to be determined using metrics from vector operations such as cosine similarity. This means that one can identify synonyms of a word by finding words whose vector representations of neighboring words are similar to those of the focal word. With a list of initial words and a collection of documents as inputs, the algorithm can expand the list to include words in the documents that are synonymous with the initial words. This capability renders \textit{word2vec} an ideal tool for constructing dictionaries to capture topics that are multi-faceted and/or evolving, and thus for which it is too difficult or too costly for experts to manually categorize all related words and phrases. Multiple studies in accounting and finance have used this promising tool for constructing dictionaries to measure various constructs, including the sentiment of financial text (Du et al. 2021), corporate culture (K. Li et al. 2021), and \textit{voluntary} HC disclosures (Zhang 2022).

As with any machine learning task, the quality of inputs is crucial for the quality of outputs (Geiger et al. 2021). In the case of \textit{word2vec}, the model should be trained on text that is known to discuss the topic for which a dictionary is being developed in order to guarantee the lexicon’s construct validity. We therefore train the model on a large collection of documents unambiguously related to HC disclosures, which is to say the actual HC disclosures made under the new regulation that were hand-collected from 3,953 unique firms’ 10-K filings.

Table 1 summarizes the sample selection process for these HC disclosures. We start with all the 10-K Forms filed with the SEC’s online filing system (i.e., EDGAR) during the first year of the 2020 regulation (i.e., from November 9, 2020 to November 8, 2021). From these 7,185 filings, we exclude 3,219 forms from filers that are not included in the Compustat and CRSP
databases. Most of these filers are private companies, trusts, closed-end funds, or partnerships that are not required to comply with the regulation, or that do not have a significant number of employees. After this exclusion, there remain 3,966 10-Ks from filers that are included in EDGAR, Compustat, and CRSP. From these we exclude five filings belonging to fiscal year 2019, three filings from firms that outsource all of their personnel (i.e., according to Compustat they do not have any employees), and two filings that do not contain HC disclosures. Finally, we exclude three duplicate filings, keeping only the first 10-K in the case of multiple filings by the same company. After this process, 3,953 unique corporate filings remain. We use the HC disclosures from these filings to train the word2vec model. Using one disclosure from each company ensures that each company is given equal weight.

Taken together, the HC disclosures of these 3,953 unique firms consist of approximately two million words. We share these disclosures and their unique company identifiers in a single text file named “JIS_Data_HC_Disclosures.txt.” The identifiers include CIKs, company names, filing dates, and fiscal reporting periods. The disclosure text and identifiers for each company are enclosed in a <DOC> tag. We provide a utility function in “JIS_HC1_util.py” for extracting the texts and associated company identifiers from the single text file and saving this data to individual text files or to a combined CSV file.

Another important input into word2vec modelling is the list of initial seed words. We next describe how we have chosen these words.

Initial seed words

To ensure that our initial seed words cover various important dimensions of HCM practices, we first identify five broad topics according to the recommendations of standard setters
and survey evidence. In its preliminary framework for HC disclosures, SASB identified three key HCM issues that are especially relevant to the long-term sustainability of a company: “Employee health and safety,” “Labor practices,” and “Employee diversity, inclusion, and engagement.”

Our first three categories address these issues. We include “Compensation and Benefits” as an additional category because compensation and benefits are the second and third most important drivers of employee satisfaction according to a survey conducted by the Society for Human Resource Management (SHRM), with the top driver being “respectful treatment of employees at all levels,” which is already included in DEI. Our last category captures employee demographics and other HCM issues.

To come up with the initial word lists for these five categories, we take a two-pronged approach. First, we examine the issues or metrics that firms should disclose as recommended by standard setters, regulators, and various interest groups such as institutional investors. Second, we manually read a large sample of HC disclosures in 10-Ks to identify the common words and phrases that firms use to discuss their HCM practices. In Table 2, we present our seed words for each of the five categories. Because we train the word2vec model on actual HC disclosures, we can choose a large number of words and phrases as seed words, and even include certain words that may appear to be too general in the context of a full 10-K document, but that are domain-specific in the context of HC disclosures. This allows us to identify words and phrases that otherwise would have been missed had a smaller number of seed words been used. For example, the words “fair” and “fairly” can take multiple different meanings in a 10-K; however, in the

---

5 Zhang (2022) provides a comprehensive summary of these recommendations in her Online Appendix 1.
context of HC disclosures, they most likely refer to whether employees receive fair treatment or are treated fairly. Including words such as these that are too general in the larger context of a 10-K, but that are specific enough in the narrow context of HC disclosures, allows us to uncover other HC-related words that may be missed when the learning algorithm is applied to entire 10-Ks as in Zhang (2022).

**Final Word lists**

We feed the seed words into word2vec to allow the model to discover similar words and phrases. Words and phrases are considered to be “similar” if they are used in the same or similar context as the seed words. For these similar words and phrases, we initially keep those with a cosine similarity score of at least 0.5, rather than keeping the top-300 most similar words for each of the seven categories as in Zhang (2022) or the top-500 words for each of the five categories as in K. Li et al. (2021).6 We make this choice for several reasons. First, keeping the top “n” words for each category assumes that there is an equal number of similar words for each category. This is not a valid assumption. Among the various dimensions of HCM, some dimensions can legitimately have more keywords than others due to their greater complexity. Second, within each category, choosing the top “n” words may give some seed words an unduly high weight simply because they happen to have more similar words with high cosine similarities.7 This can reduce the comprehensiveness of the final word list. By choosing a threshold based on cosine similarity, we do not force each category to have the same initial weight and we do not allow certain seed

---

6 With word embeddings, the cosine similarity score is bounded by negative one and positive one. A score of positive one indicates that two words always occur in the same context. A negative score indicates that the two words rarely occur in the same context and they are possibly antonyms.

7 If seed word “A” has a large number (say, more than 300) of similar words with high similarity scores (say, all above 0.80), whereas seed word “B” has many words with similarity scores that are well above the reasonable 0.5 threshold but less than 0.8 (but some of them are very close to 0.8), then choosing only the top “300” words would exclude all the similar words of seed word “B” from being considered. This would result in an overrepresentation of words that are similar to seed word “A” and an underrepresentation of words that are similar to seed word “B.”
words to dominate the list. Third, prior studies show that using a cosine similarity threshold of 0.5 yields stronger results. For example, Kee (2019) proposes a new approach for identifying peer firms using \textit{word2vec} and finds the new approach outperforms existing industry classification systems when the threshold for the cosine similarity is greater than 0.5. Erfani, Cui, and Cavanaugh (2021) use \textit{word2vec} to identify similar risks across different projects based on risk descriptions and find that two risks are meaningfully related to each other when the cosine similarity is above 0.5.

To ensure that our word list captures antonyms, we also include words in our initial word list if they have a highly negative cosine similarity (i.e., cosine \leq -0.5), consistent with the notion that even antonyms can appear in a similar context.\footnote{We thank an anonymous reviewer for this insightful suggestion.} For example, companies may use “bias” or “anti-bias”, two semantically opposite words, when they discuss their DEI policies. To investigate the extent of such words, in Figure 1 we plot the distribution of cosine similarity scores of all words with the seed words. The positively skewed distribution indicates that more words have a positive cosine similarity score, while the left tail of the distribution indicates that there is only a small percentage of words that have a high negative cosine similarity score.

At the cutoff of an absolute value of 0.5, our initial word list consists of 7,018 keywords. This large list of initial similar words ensures the comprehensiveness of our final word list. We take three steps to screen and classify these words. As the first step, we manually screen each of these keywords to make sure that they relate to HCM. Next, we assign a term to the category of the related seed word. When a term is similar to seed words from multiple categories (e.g., “DEI” and “Health and Safety”), we assign the term to the category for which the term has the highest
average similarity score with the category’s seed words. As the last step, we further review the classification of all the terms and manually re-classify them as appropriate. These procedures are standard with a word2vec implementation, and serve to improve the quality of the word lists. Both K. Li et al. (2021) and Zhang (2022) screen out some words that do not pass the face validity check. Some of the words that we drop are not related to HCM, while others are related to HCM but may be too general for documents covering many topics other than HCM (e.g., “fair”). We exclude such words so that our final word list will have broader applicability to texts beyond the confirmed HC-specific disclosures used to construct our lexicon.

Table 3 presents the final word lists for each of the five categories. The keywords consist of seed words and similar words identified by the algorithm, for a total of 1,285 terms. Each of our word lists captures an important aspect of a firm’s HCM practice, and collectively the lists enable the construction of measures that capture the firm’s overall HC disclosure. Some of the keywords are acronyms. We provide their full spellings in Table 4. For the “Health and Safety” category, we further classify keywords into those related to COVID-19 (70 terms) and those that are more generic in nature (157 terms). The keywords related to COVID-19 are useful for examining questions related to firms’ HC-related responses to the pandemic, but are less likely to be of interest to future researchers working outside of that particular context.

Following the prior literature, we classify the keywords into five categories largely based upon the keywords’ similarity with the initial seed words for each category (i.e., computationally determined, but subject to a manual review as described above). To assess the appropriateness of this classification, we provide a visualization of the clusters of keywords that derive from our classification, and we compare the results to those generated by an automated classification using
the K-Means clustering algorithm. To visualize our classification, we extract the vectors of the keywords from the word2vec model, which represents each word using a vector of 100 numbers. We then reduce the dimensionality of the data using Principal Component Analysis (PCA) to two and three dimensions, respectively, by keeping the first two and three components. Because we further break down keywords related to health and safety into those specific to COVID-19 and those that are more generic in nature, our classification consists of six categories (or clusters). For the automatic clustering, we use the K-Means clustering algorithm to classify these keywords into six clusters for easier comparison.

Panels A and B of Figure 2 present a 2-D visualization of our classification and the K-Means clustering classification, respectively. Each dot represents a keyword and the color of the dot represents its category or cluster. As shown in Panel A, the clustering of the keywords indicates that our classification works well. For example, the keywords for “DEI” and those for “Compensation and Benefits” cluster together coherently in the graph. In Panel B, the K-Means clustering may appear at first blush to be doing a better job. However, the largest K-Means cluster consists of more than 80% of all keywords, whereas some of the K-Means clusters consist of a much smaller number of words. Manual inspection of the larger cluster suggests that many of the keywords could have been further classified into finer clusters. Manual review of one of the smaller clusters shows that it consists of the following short list of words: “age, background, black, disability, discrimination, equality, ethnic_minority, ethnically_diverse, ethnicity, female, gender, gender_identity, harassment, hispanic, latino, latinx, lgbtq+, male, marital, minority, orientation, race, racial, religion, representation, self-identified, sex, sexual, veteran, veteran_status, woman.” Although these words all clearly relate to DEI, they represent only a subset of the DEI keywords.

---

9 We thank the anonymous reviewer for this suggestion.
included in our word list. Figure 3 presents the visualization in 3-D space. This graph similarly indicates that our classification works well. For example, in Panel A of Figure 3, the keywords related to “Health and Safety (COVID)” from our classification cluster together neatly in a manner that was less apparent in the previous 2-D visualization. On visual inspection, Panel B of Figure 3 for the automated clustering depicts a pattern similar to that in the previous 2-D graph (i.e., Panel B of Figure 2), which is not surprising given the dominance of the one very large category that derives from this clustering process.

The previous comparison uses six clusters for K-Means clustering to correspond with the six clusters identified from our manual categorization. To alternatively find the optimal number of clusters, we calculate the silhouette scores for different numbers of clusters, ranging from N=2 to N=25. The theoretical value of the silhouette score (or coefficient) ranges from -1 to +1, with higher values indicating more coherent clusters. Figure 4 plots the silhouette scores against the number of clusters. As shown, the score peaks at four clusters, and it has the second highest value at six clusters. Given that most of the prominent HC disclosure frameworks (e.g., GRI standards for HCM, IOS standards for HC reporting, and BlackRock’s guidelines on HCM for investee companies)\(^\text{10}\) identify more than four dimensions to corporate HCM, we consider the second peak, at six clusters, to be more appropriate for automated clustering. Investigating higher numbers of clusters, we observe, for example, that at 10 clusters, the largest cluster still consists of more than 70% of all keywords even though the smallest cluster contains only 15 keywords. At 20 clusters, the largest cluster consists of 63% of all keywords, whereas some of the smallest clusters contain fewer than 10 keywords. These additional analyses suggest that automated clustering has its

limitations as this process is not able to distinguish the multidimensionality of human capital management to a satisfactory level. Finally, a further important insight offered by these analyses is that the algorithm’s classification of most keywords into a single cluster indicates that our keyword list captures well the single broad underlying construct of human capital management.

The less-than-satisfactory results from automatic clustering could also be due to the fact that word embeddings generated by word2vec are context-independent. This means that a word has the same vector representation regardless of the context in which it is used. Due to this inherent limitation, word2vec is not able to effectively capture polysemy (i.e., words or phrases that can have different meanings depending on the context). Nonetheless, this negative impact should be relatively small and does not undermine the validity of our methodology because we train the word2vec model on a corpus of confirmed HC disclosures. The meanings of words used by companies to describe their HCM practices tend to be less varied across HC contexts compared to their use in general language. An alternative approach for refining the classification of HC disclosures into different topics is to use machine learning. We discuss more about this approach in the next section.

IV. APPLICATIONS INVOLVING OUR HC LEXICON AND DATA

An Application of Our Lexicon

To illustrate the application of our lexicon, we use it to examine the trend of HC disclosures in proxy statements. Public companies are required to provide proxy statements to shareholders before annual meetings that include information about compensation for executives and key employees. We construct a few measures using our lexicon for HC disclosures in proxy statements over the period from 1994 (the starting year of electronic filing) to 2022. These measures include
a count of HC keywords and the percentage of HC keywords relative to the total number of words in the proxy statement. We construct measures for each of the five categories, as well as an aggregate measure that is computed as the sum of the five categorical measures. Technical details and the corresponding Python code underlying the measures’ construction are provided in the Appendix.

In Figure 5, we plot the volume of HC disclosures in proxy statements from the period of 1994 to 2022. The left vertical axis shows the absolute HC keyword count and the right axis tracks the HC keyword count as a percentage of the total number of words in the proxy statement. Over the 1994 to 2003 time period, there are about 300 HC keywords in an average proxy statement. The absolute word count increases gradually to 400 words over the 2004 to 2006 period, followed by a jump to more than 600 words in 2007. The change in 2007 was due to a new SEC regulation that required firms to provide expanded disclosures about executive and director compensation. Our lexicon successfully captures this structural change in the reporting environment. In 2010 and 2011, there is another large increase in the HC keyword count relative to 2009, which is due to the SEC’s Proxy Disclosure Enhancements (effective February 28, 2010) that amended Regulation S-K Items 401, 402 and 407. Our lexicon again exhibits construct validity as it captures this structural change in regulatory reporting requirements. After 2011, the keyword count remains stable at more than 800 words, except for a dip in 2018 that was caused by a greater number of firms filing proxy statements in that year, with most of the new filers being smaller companies. In absolute terms, the amount of HC disclosure is fairly stable over the 1994 to 2006 period, with the percentage of

---

11 Since this is for illustrative purposes, we make only a minor adaptation to our lexicon that involves excluding the “health and safety” keywords related to COVID-19 as these are obviously not relevant to the longer time series of documents used in our illustration.
keywords hovering at around 2.5%. In 2007, the percentage increases to approximately 3% and varies very little over the years after that.

In Figure 6, we break down the percentage of keywords by category. The category with the largest number of disclosures is “Compensation & Benefits,” followed by “Demographics and Other,” and “Labor Relations & Culture.” The two categories with the lowest number of disclosures are “DEI” and “Health & Safety (General).” This ranking is consistent with the contents of proxy statements, which provide a large amount of information about compensation, benefits and demographics, especially for executives and directors. Turning to the trend, there is an obvious jump in 2007 for disclosures in “Compensation & Benefits” and “Demographics & Other,” due to the SEC’s regulation change in that year. Even though the number of disclosures in the “DEI” and “Health & Safety” categories is low over the sample period, there is an evident increase in 2020 and 2021, consistent with both the growing awareness of the importance of DEI in recent years, as well as the many employee health and safety concerns that arose during the COVID-19 pandemic.

In summary, this simple example illustrating the application of our lexicon to corporate proxy statements (i.e., “out-of-sample” documents) provides substantial evidence of the construct validity of our HC dictionary as the documented trends map into well-known regulatory and societal changes over the period under study.

An Application Using Our Textual Data

The advancement of Large Language Models (LLMs) underscores the growing importance of textual data, and particularly labeled textual data, because labeled textual data can be used to improve model performance through pre-training or fine-tuning. For example, users can leverage
our HC disclosure text to fine-tune the base BERT model, thereby enabling BERT to identify HC disclosures and, even further, to classify the disclosures into specific topics. To showcase this potential, we use the confirmed HC disclosures underlying our study as an annotated dataset, supplemented by non-HC disclosures in Item1 of 10-Ks. We extract 83,961 sentences from the HC disclosures and generate a random sample of twice as many non-HC sentences. Following the standard practice in machine learning, we use 80% of these sentences as the training set and 20% as the test set. We use “bert-base-uncased” as the base model and fine-tune it for a binary classification task (i.e., HC vs non-HC). We fine-tune the model on a single powerful NVIDIA A100 GPU on Google Colab Pro. Evaluation on the test set indicates an accuracy score of 97.46%, precision of 96.05%, recall of 96.33%, and F1 score of 96.19%. These results suggest that the fine-tuned model performs very well in out-of-sample prediction.

We next apply the fine-tuned model to 10-Q disclosures to see whether it continues to perform well on disclosure text that is different from the training set (i.e., contents in 10-Qs tend to be different from those in the 10-Ks that are the source of the training data). For this exercise, we manually label 1,000 sentences, 437 of which are confirmed HC sentences. We use the fine-tuned model to predict whether a sentence is related to HC. The output from the model is a probability, ranging from 0 to 1, assessing the likelihood that the textual input belongs to the positive class (i.e., is an HC sentence). Because the model does not assign a label, the user is responsible for determining the appropriate threshold above which a sentence is to be labeled as an HC sentence. While 50% may seem a convenient choice as the threshold for binary

---

12 For a clear and accessible demonstration on how to fine-tune a BERT model, see https://github.com/PradipNichite/Youtube-Tutorials/blob/main/FineTune_BERT_Model_Youtube.ipynb.
13 The hyperparameters are: epochs = 2, batch_size = 16, learning_rate = 2e-05, and warmup_steps = 10000.
14 In a binary classification task, “accuracy” is calculated as the number of correct predictions / the total number of predictions made by the model, “precision” is (True Positives)/(True Positives + False Positives), “recall” is (True Positives)/(True Positives + False Negatives), and “F1 score” is the harmonic mean of precision and recall.
classification, this choice is often not optimal. To find the optimal choice, we vary the threshold from 0 to 1 at a step of 0.01 and plot the precision, recall, and F1 score for each threshold in Figure 7. For this illustration, we choose the threshold that results in the highest F1 score, which corresponds with a probability threshold of 86% - i.e., a sentence is labelled as HC-related only if the probability predicted by the model for that sentence is 86% or higher. At this threshold, the corresponding precision, recall, and F1 score are 85.16%, 85.35%, and 85.26%, respectively, for the HC class. These metrics indicate that the fine-tuned model performs impressively well on out-of-sample classification.

To provide a more complete picture of the model’s performance, we present the confusion matrix and classification report in Figures 8 and 9, respectively. The confusion matrix shows the number of instances that the model predicted correctly (incorrectly) on (off) the diagonal. For example, the second row of the matrix indicates that the model correctly (incorrectly) predicts 373 (64) out of the 437 confirmed HC sentences. The classification report provides three metrics (precision, recall, and F1-score) for each class as well as their averages over the two classes and the number of instances (referred to as “support”) for each class. As shown in Figure 9, the HC class has a precision, recall, and F1-score of 85.16%, 85.35%, and 85.26%, respectively, consistent with what we reported earlier. The Non-HC class has a slightly higher value for each of these metrics, indicating that the model does a marginally better job at identifying non-HC sentences. Both the weighted average and the macro average provide a summary score over the two classes for each of the three metrics, except that the weighted average gives a higher weight to the class with more instances (non-HC sentences in this case). Because the evaluation dataset is relatively

---

15 For a guide on how to implement this automatic selection process, see https://ploomber.io/blog/threshold/.
balanced between the two classes, there is little difference between the two averages in this particular example.

The preceding illustration demonstrates the use of our data for fine-tuning a BERT model for binary classification of HC disclosures. Further classification of HC content into more granular topics can also be done using a BERT model or another LLM. One way to approach this task is to first generate word embeddings for each HC sentence by using a fine-tuned model or another pre-trained model, and then to group similar sentences together based on their word embeddings using an automatic clustering algorithm.

**Keyword- Versus Machine-learning-based Classification**

In the preceding subsections, we have illustrated how to use our lexicon and data for identifying and classifying HC disclosures using two different approaches, keyword-based vs machine-learning-based. Each of these methods has its own advantages and limitations, which we briefly discuss in this section.

Keyword-based classification may require pre-processing for normalizing certain words such as lemmatization and phrasing.¹⁶ These additional steps may introduce noise, bias, or errors in the form of incorrect term reduction, loss of important contextual information, or incorrect parsing of complex phrases. Moreover, the keyword-based approach provides a deterministic result based upon the presence or absence of certain keywords and may not capture the context and other nuances of the language. Nonetheless, the keyword-based approach is transparent and easy to use. Such transparency allows fellow researchers to easily understand the classification

---

¹⁶ An alternative approach is to employ regular expressions to account for various forms of a word or phrase. However, this method can become cumbersome, particularly when dealing with a large number of keywords.
process and also allows them to better evaluate, replicate, and potentially build on the prior research. The inherent simplicity of this approach makes it an accessible tool for researchers across various levels of expertise.

On the other hand, an ML-based approach, especially if it leverages the state-of-the-art LLMs, may provide better performance due to the ability of these models to understand complex linguistic and semantic features. However, ML models may need to be pre-trained on domain-specific text or fine-tuned on annotated datasets for improved performance. This requires a high level of technical expertise. The utilization of ML-based approaches may also be constrained by the availability of high-quality, domain-specific text or by the tremendous effort involved in creating a labeled dataset. Ultimately, the decision regarding which method to use should be made based upon the nature of the research question, the characteristics of the text under analysis, and any other constraints that the researcher may face, including their own technical expertise.

In summary, each method has its own strengths and weaknesses. Keyword-based classification is effective and efficient when the categories are well-defined and there are clear lexical cues that distinguish them. On the other hand, ML-based text classification becomes more appropriate when the categories are more nuanced and there are no obvious keyword cues. In either case, care should be exercised in its implementation. When researchers adopt a keyword-based method that is reliant on lexicons created by other researchers, it is advisable to adapt the keyword to suit the characteristics of the adopters’ textual corpus and research objectives. This is akin to adapting a pre-trained machine learning model to the specific needs of a project through fine-tuning. When researchers choose an ML-based approach, it is crucial to fully document and disclose the specifics of the model and the training process. Such a practice helps reduce the opacity commonly linked to machine learning techniques.

Electronic copy available at: https://ssrn.com/abstract=4197489
V. FUTURE RESEARCH OPPORTUNITIES

As the world economy continues to be predominantly knowledge-based and the social pillar of ESG becomes increasingly important (Drei et al. 2019; CPA Canada 2023), there is a growing interest on the part of many stakeholder groups related to corporate HCM practices. In this section, we suggest some ideas that researchers in accounting and finance may pursue using our lexicon. As discussed in the literature review, prior studies examining the relation between HCM practices and firm value tend to use measures from external sources. There is a great opportunity to construct measures from firm-provided disclosures to gauge HCM practices for a large cross-section of firms. Using these measures as a complement to, or substitute for, existing measures, researchers may conduct more powerful tests on the relation between HCM practices and firm performance. Moreover, our lexicon covers five major categories, allowing for the construction of more granular measures of HCM practices. With these finer measures, researchers can study which employee treatment practices contribute more to value creation or to other important performance outcomes such as innovation. Answers to such questions have significant practical implications.

Firms can provide HC disclosures through many channels, including their annual reports, quarterly reports, proxy statements, earnings conference calls, and sustainability reports. Beattie and Smith (2010) surveyed HR directors of UK listed companies and found that managers consider annual reports to be the most effective venue for disclosing a firm’s HC policy and practices to external parties. Since that time, however, there has been a dramatic increase in corporate issuances of stand-alone sustainability reports, and most companies offer a significant amount of sustainability-related information on their websites. There is a notable dearth of large-sample empirical evidence related to the determinants and consequences of companies’ choice of HC
disclosure channel(s). There are rich research opportunities to examine what, where, and why firms provide particular HC disclosures, and what the implications of these choices are to the firm’s various stakeholders.

Among alternative disclosure channels, earnings conference calls are noteworthy. As a special form of corporate voluntary disclosure, conference calls offer an opportunity for interaction between corporate executives and financial analysts. The Q&A portion of conference calls, which is more extemporaneous (Chen, Demers, and Lev 2018), has been shown to elicit informative disclosures (Matsumoto, Pronk, and Roelofsen 2011). The extent to which managers volunteer, or analysts probe for, details about the firm’s HCM is likely to be indicative of the materiality of HC to the firm’s performance and prospects. Multiple studies develop measures using voluntary disclosures from earnings conference call transcripts. For example, Hassan et al. (2019) develop a firm-level political risk measure, Sautner et al. (2023) provide measures of firms’ climate change exposures, and K. Li et al. (2021) create a dictionary for measuring corporate culture from the Q&A portion of earnings calls. Future research could similarly develop a measure of the firm’s HC-centricity or sensitivity, and explore how these measures influence capital market activity or analyst forecast updates, particularly in response to potentially HC-sensitive events such as social movements or changes in HC-related legislation.

Our word lists can be used to extract HC disclosures from conference call transcripts or other forms of corporate communications, and the extracted text can then form the basis from which other disclosure attributes such as readability (F. Li 2008), tone (X. Huang, Teoh, and Zhang 2014), specificity (Hope, Hu, and Lu 2016), numerical intensity (Henry 2008), forward-looking information (Muslu et al. 2014), and similarity (Brown and Tucker 2011) could be measured. These measures, in turn, can be used to gain a more nuanced understanding of a firm’s HC.
disclosure decisions. For example, an overly optimistic tone that is not justified by real performance may indicate that a firm is “social washing.” To the best of our knowledge, there is little evidence related to the extent, determinants, and consequences of social washing in HC disclosures, even though greenwashing in sustainability reporting has become an issue of top priority for regulators around the world.¹⁷

An additional opportunity available to researchers interested in the technical aspects or methodology of textual analysis is to pre-train or fine-tune a BERT model or another LLM on an extensive corpus of HC disclosure text. In our illustration, we have demonstrated this potential by obtaining satisfactory results through a modest fine-tuning process on a relatively small dataset, coupled with limited hyperparameter tuning. Those who plan to exploit this opportunity can incorporate our data into their training dataset and leverage our lexicon for identifying additional documents that are rich in HC-related content.

VI. CONCLUSION

Human capital is frequently a company’s most valuable asset, yet unlike other important asset classes, HC is not subject to well-defined measurement or disclosure rules. The construct of corporate human capital is itself so nebulous that the SEC actually refused to define it in the new disclosure regulations (Bourveau et al. 2022). We use a semi-supervised machine learning algorithm (word2vec) that is trained on a confirmed set of recent corporate HCM disclosures to develop an HC-related lexicon consisting of 1,285 terms classified into five disclosure categories: diversity, equity, and inclusion; health and safety; labor relations and culture; compensation and

¹⁷ For example, the SEC launched the Climate and ESG Task Force within the Division of Enforcement in March 2021, with the mandate to identify and investigate potential corporate misconduct related to climate and ESG issues. (https://www.sec.gov/securities-topics/enforcement-task-force-focused-climate-esg-issues).
benefits; and demographics and other. We share our dictionary in a machine-readable format, together with the textual data and Python code used to construct it. We also provide an example of its application to corporate proxy statements. Researchers can modify our code to construct their own lexicon related to another topic of interest, or use our HC-related lexicon to pursue the many research opportunities that are available in this burgeoning area of study. In addition, we demonstrate the potential of using our textual data to fine-tune a LLM as an alternative approach for identifying and classifying HC disclosures.
APPENDIX – PYTHON CODE

In this appendix, we present the Python code for using our lexicon.\textsuperscript{18} We also provide the code for training a \textit{word2vec} model in the event that researchers want to expand our HC word list or create an entirely new word list to capture a different topic. These codes have been tested on a computer running Windows 11. Please see the “requirements.txt” for the list of Python libraries required for running the codes.

These codes are designed to be memory efficient so that researchers can use them to process collections of documents that are too large to be loaded into the memory all at once. Moreover, our code supports parallel processing by using multiple CPU cores to cut the processing time. For example, when we generate the variables for the more than 186,000 proxy statements over 1994-2022, it takes only about 20 hours using ten logical CPU cores. These proxy statements contain 3.3 billion words and occupy a disk space of more than 20GB. Using 10 workers, it takes about 38 hours to train a new \textit{word2vec} model on this same large corpus of documents. Most of the time (28 hours) is spent on preprocessing the documents, which includes training the phrase model.

\textbf{Code for generating keyword-based measures using our lexicon}

In what follows, we provide some technical guidance on how to apply our lexicon to create keyword-based measures. The keywords presented in Table 3 are lemmatized and contain phrases.\textsuperscript{18} Although our dictionary is intended for general purpose use, as suggested by Bochkay et al. (2022), researchers are encouraged to make their own adaptations to our lexicon to tailor it to their particular setting. In information retrieval or classification, a proper balance needs to be struck between “precision” and “recall,” which are analogous to Type I and Type II errors. In the current context, this would require that a researcher examine the sample of results generated by our lexicon, and then make whatever changes may be necessary to the keywords to suit their specific research context (e.g., excluding certain keywords, including new keywords, and/or modifying the forms of certain keywords).
There are two ways to use keywords in this format. First, one can rewrite each keyword into a regular expression to match the keyword and its alternative forms. Re-writing keywords into regular expressions can be cumbersome for hundreds of keywords. An alternative and more efficient approach is to tokenize the text, lemmatize the words, join certain words into phrases, and then check each word or phrase to see whether it is on the keyword list. Our code takes this second approach.

For greater readability and ease of use, we group our code into six modules (i.e., individual source code files). We briefly describe these modules below. More notes can be found in the code files.

(1) JIS_HC0_config.py

This module allows users to define parameters, options, and settings for the project – e.g., where the keyword list is located, where the source documents can be found, how many parallel processes to run, where to save preprocessed files and final outputs, and what parameters to use for phrase model training and word2vec training.

(2) JIS_HC1_util.py

This model contains functions for performing certain routine tasks such as generating a list of all files in a directory, reading text from a file, tokenizing a text into sentences, lemmatizing words, and outputting the results. Users may find these functions useful for many other NLP projects.

---

19 For example, “equality” may appear in a plural form or in upper case, or have one or more of its letters capitalized. To capture these variants, one needs to write a regular expression as “(?i)equality|ties\?”. To make the whole regular expression ignore case, a flag (e.g., “re.I” for Python) can be passed to the regex engine instead of using the inline flag, “(?i)”. 

---
(3) JIS_HC2_load_keywords.py

This module loads HC keywords from the CSV file and generates three objects for use in later steps: a list of all HC keywords, a list of all HC keywords that are phrases (bigrams or trigrams), and a Python data structure known as a “dictionary” for storing HC keywords for each of the five categories.

(4) JIS_HC3_generate_keyword_variables.py

This module generates variables using our HC keyword lists. Users do not need to run the previous three modules separately. They are imported into this module automatically when this module is run. This module defines some additional functions. It performs the following steps by using these additional functions as well as the functions or objects from the previous three modules:

1. Load keywords from the CSV file. This is done by importing the three objects from “JIS_HC1_util.py.” Users do not need to run “JIS_HC1_util.py” separately.
2. Traverse the specified directory (including sub-directories) to find all txt files that need to be processed.
3. Open each txt file and extract the text.
4. Tokenize the text into sentences and further tokenize each sentence into words.
5. Convert all words into lower cases and lemmatize them.
6. Join words that are part of a phrase with an underline “_”, according to the list of phrases in our keyword list. For example, “affirmative action” is converted to “affirmative_action”.
7. Count the frequencies for all keywords in each category.
(8) Identify sentences containing any of the keywords and count the number of such sentences as well as the word count of all such sentences.

(9) Get the word count and sentence count of the entire document.

(10) Output the txt file name (i.e., the document identifier) and variables as a new row in a CSV file.

Steps 3 through 10 are performed in parallel. Users can specify the number of workers in “JIS_HC0_config.py”. By default, this module uses all available logical CPU cores, less two.

For tokenization, we mostly use tools from NLTK, which is known for its flexibility, a large support community, and great performance for this task.\textsuperscript{20} For lemmatization, we provide two options, NLTK and Spacy.\textsuperscript{21} The default is the “WordNetLemmatizer” available with NLTK. Users can choose to use the Spacy lemmatizer by changing the setting in “JIS_HC0_config.py.” Although the Spacy lemmatizer is slightly quicker, users should be aware that the Spacy lemmatizer converts all pronouns to “-PRON-”. In addition, the Spacy lemmatizer keeps proper pronouns such as “Native Americans” unchanged (i.e., instead of converting it to “Native American”). To use the Spacy lemmatizer, some keywords may need to be adapted.

This module takes a list of text files from the specified directory as inputs. It is ideal for the use case where one works with a large corpus of documents, for example, 10-K or 10-Q filings, each of which is saved in a separate txt file. Users can try out the code by using cleaned 10-K filings available at https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/. The code can be easily adapted for use in cases where the text is stored in another format, for example, in CSV files.

\textsuperscript{20} https://www.nltk.org/
\textsuperscript{21} https://pypi.org/project/spacy/
The next few modules are for training a *word2vec* model.

**Code for training a *word2vec* model**

(5) *JIS_HC4_train_phrases.py*

This module trains a phrase model using Gensim, a popular Python library for NLP. As part of the preprocessing for *word2vec* training, we use the phrase model to generate phrases so that the *word2vec* algorithm can identify not only single words but also phrases that are similar to the seed words. This module can be used for many NLP projects that benefit from phrase training, such as LDA topic modelling.

Training a phrase model is a time-consuming process for a large corpus because the algorithm needs to go through the entire corpus multiple times. To reduce the processing time, our code saves pre-processed sentences to a folder specified by the user in “JIS_HC0_config.py.” It then loads the pre-processed sentences from the folder when they are needed again in later steps. To discover both two-word phrases (bigrams) and three-word phrases (trigrams), this module performs the following steps:

(a) Preprocess the text (by tokenizing the text into sentences, tokenizing the sentences into words, lemmatizing the words, and further normalizing the words), and save these sentences consisting of unigrams (i.e., single words, without phrases being tagged).

For the normalization procedure, we convert words containing numbers to “1” and convert words containing punctuations (excluding “-” and “_”) to “,”. This can reduce the vocabulary size and speed up the later training process. Note that we do not simply drop

---

22 [https://pypi.org/project/gensim/](https://pypi.org/project/gensim/)
23 If any word containing punctuations or numbers is known to be important (e.g., “COVID-19”), it should be added to the whitelist.
non-alphabetic words, punctuations, or stop words, because doing so will break the flow of a sentence and result in phrases consisting of words that are separated by a number, punctuation, or stop word in a sentence.

(b) Load the preprocessed unigram sentences from the disk and feed them into the algorithm to train a bigram model.

(c) Obtain all bigrams discovered by the model and generate a list of valid two-word phrases (bigrams) by removing those containing stop words, numbers, punctuations (other than “_” and “-”) or single-letter words.\(^\text{24}\)

(d) Process the unigram sentences to form phrases using the bigram model (e.g., converting “workplace safety” to “workplace_safety”), break the phrases not in the list of valid phrases from (c), and save the results (bigram sentences) to the disk.

(e) Load the bigram sentences and feed them into the algorithm again to train a trigram model.

Save the model to the disk.

(f) Obtain all the bigrams and trigrams identified by the model and generate a list of valid two-word phrases and three-word phrases by excluding those bigrams/trigrams containing stop words, numbers, punctuations (other than “_” and “-”) or single-letter words. Export the list of bigrams and trigrams into a CSV file for inspection and evaluation.

(6) JIS_HC5_preprocess_for_word2vec.py

This module generates the final inputs for training a word2vec model. The final inputs are sentences that are tokenized, lemmatized, phrased, and have all stop words, numbers, and

\(^{24}\) If a phrase containing these elements is known to be important, it should be added to the whitelist.
punctuations removed. Building on the previous phrase modelling process, it performs the following steps:

(a) Load the bigram sentences, the trigram model, and the list of valid phrases.
(b) Form phrases using the trigram model and break those not in the valid phrase list.\(^{25}\)
(c) Remove stop words, one-letter words, and words containing digits or punctuations (excluding “_” and “-”).
(d) Save the results (i.e., trigram sentences) to the disk.

The module will first look for preprocessed trigram sentences in the folder specified in “JIS_HC0_config.py.” If they are not found, it will generate these inputs by going through the whole process. In this way, users do not need to generate the inputs again when they try out new parameters for the \textit{word2vec} model. However, if users want to use new parameters for phrase modelling, they should delete the previously saved trigram sentences and bigram/trigram models so that new ones will be generated to reflect the desired changes.

(7) \textit{JIS_HC6_train_word2vec.py}

This module trains a \textit{word2vec} model, saves the model, and outputs the similar words of a given list of words into a CSV file. It performs the following steps:

(a) Load the preprocessed trigram sentences from the specified folder. If no files are found, it will generate these sentences using the txt files in the specified folder by calling related functions from “\textit{JIS_HC4_train_phrases.py}” and “\textit{JIS_HC5_final_preprocess_for_word2vec.py}.”

\(^{25}\) If one of the two words is a two-word phrase, this will result in a three-word phrase (e.g., “workplace\_safety\_program”). If both words are two-word phrases, then this will result in an occasional four-word phrase.
Users do not need to run these two modules (i.e., “JIS_HC4” and “JIS_HC5”) before they run the current module. However, users are recommended to first run the phrase training module (“JIS_HC4_train_phrases.py”) and inspect the results. Based on the results, users can fine-tune the parameters and determine whether to exclude certain phrases discovered by the algorithm. Once the decision is made, users can remove the preprocessed files and saved models, and re-run this module so that a new model will be trained using new parameters.

(b) Train a \textit{word2vec} model using Gensim based on the parameters from “JIS_HC0_config.py” and save the model to the specified folder.

(c) Load the seed words from the CSV file specified in “JIS_HC0_config.py.” Each column of the CSV file should contain the seed words for one category. The top row of each column should contain the category name (e.g., “Compensation&Benefits”). If a seed word is a phrase, it should be separated by spaces, for example, “human resource”, instead of “human_resource.”

(d) Generate similar words and export them to a CSV file.

The CSV file contains the following columns: “Category,” “SeedWord,” “SeedWordFreq,” “SimWord,” “SimScore,” “SimWordFreq,” “SimScore_ABS,” “AvgCatSim_ABS.” “SeedWordFreq” represents the number of times the seed word appears in the training corpus. “SimScore” represents the similarity score between a seed word and a similar word (“SimWord”), with a higher value indicating greater similarity. “SimWordFreq” represents the number of times that the similar word appears in the corpus. “SimScore_ABS” is the absolute value of “SimScore,” and is useful for considering similar words with both positive and negative similarity scores.
“AvgCatSim_ABS” is the average absolute similarity score of a word with all the seed words in a category. When a word is found to be highly similar to seedwords from multiple categories, the word is assigned to the category with the highest “AvgCatSim_ABS.”

For each seed word, users can choose how many similar words to keep, for example, the top 300 words or top 500 words with the highest similarity scores. Alternatively, researchers can choose to keep similar words whose similarity scores meet a certain threshold, such as +0.5 or an absolute value of 0.5. Users can change the settings in “JIS_HC0_config.py.” Under either approach, a word may be found to be highly similar to multiple seed words that belong to different categories. This gives rise to a situation where a word can be classified into multiple categories. One way to avoid this is to compute the average similarity score between the word and seed words in each category and re-assign the word to the category having the highest average score. The program outputs another list of similar words reclassified in this way to a CSV file. Ultimately, however, the researcher should determine which category or categories the word belongs to.
REFERENCES


Accounting Standards and Regulation, 45 (2): 221–47. 


doi:10.1080/1350485032000100297.


FIGURE 1 Distribution of Cosine Similarity Scores

This graph plots the distribution of the cosine similarity scores between the seed words and other words in the corpus of HC disclosures used for training the word2vec model.
FIGURE 2 Two-Dimensional Visualization of Keyword Classification

Panel A: Our classification

This graph visualizes our classification of keywords in a two-dimensional space. Each dot represents a keyword, with its color indicating the respective category of the keyword. The horizontal and vertical axes capture the first two components from PCA analysis.
FIGURE 2 Two-Dimensional Visualization of Keyword Classification

Panel B: Automated K-Means Clustering

This graph visualizes the classification of keywords from K-Means clustering in a two-dimensional space. Each dot represents a keyword, with its color indicating the respective cluster of the keyword. The horizontal and vertical axes capture the first two components from PCA analysis.
FIGURE 3 Three-Dimensional Visualization of Keyword Classification

Panel A: Our classification

This graph visualizes our classification of keywords in a three-dimensional space. Each dot represents a keyword, with its color indicating the respective category of the keyword. The three axes capture the first three components from PCA analysis.
FIGURE 3 Three-Dimensional Visualization of Keyword Classification

Panel B: Automated K-Means Clustering

This graph visualizes the classification of keywords from K-Means clustering in a three-dimensional space. Each dot represents a keyword, with its color indicating the respective cluster of the keyword. The three axes capture the first three components from PCA analysis.
FIGURE 4 Silhouette Analysis for Optimal Number of Clusters

This graph plots the number of clusters and their corresponding Silhouette scores from K-Means clustering.
The left vertical axis shows the absolute HC keyword count. The right vertical axis tracks the relative HC keyword count as a percentage of the total word count of an average proxy statement in a given year.
FIGURE 6 Trend of HC Disclosures in Proxy Statements by Category over 1994-2022

This figure shows the relative HC word count of each category as a percentage of the total word count of an average proxy statement in a given year.
This figure shows three key classification metrics (precision, recall, and F1-score) across a range of probability thresholds from 0 to 1. The left vertical axis represents the values of these metrics. The right vertical axis depicts the number of sentences labelled as HC-related for each corresponding threshold.
This figure presents the confusion matrix of the binary classification of sentences into HC and non-HC. The left axis denotes the true labels assigned through human labeling, while the right axis represents the labels assigned based upon the probabilities generated by the fine-tuned BERT model. The matrix distinguishes between true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications. As an illustration, the top-left quadrant of the matrix quantifies the number of TNs - i.e., instances where true non-HC sentences were correctly identified by the model.
This figure presents the classification report of the binary classification of sentences into HC and non-HC. The report provides a comprehensive overview of the model’s performance, including precision, recall, F1-score, and support for both classes. Precision indicates the accuracy of positive predictions, recall measures the model’s ability to capture all relevant instances, and the F1-score is the harmonic mean of precision and recall. The support metric indicates the number of instances for each class.

Electronic copy available at: https://ssrn.com/abstract=4197489
### TABLE 1 Sample Selection

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All 10-K forms filed during the first year from November 9, 2020 to November 08, 2021</td>
<td>7,185</td>
</tr>
<tr>
<td>Less:</td>
<td></td>
</tr>
<tr>
<td>Forms of filers not covered by Compustat and CRSP</td>
<td>(3219)</td>
</tr>
<tr>
<td>Number of 10-Ks in the intersection of EDGAR, Compustat, and CRSP</td>
<td>3,966</td>
</tr>
<tr>
<td>Less:</td>
<td></td>
</tr>
<tr>
<td>Filings for fiscal year 2019</td>
<td>(5)</td>
</tr>
<tr>
<td>Forms of firms having no employees</td>
<td>(3)</td>
</tr>
<tr>
<td>Filings that do not contain HCM disclosures</td>
<td>(2)</td>
</tr>
<tr>
<td>Duplicate filings from the same filer (keeping the first one)</td>
<td>(3)</td>
</tr>
<tr>
<td>Number of HC disclosures from the same number of unique firms</td>
<td>3,953</td>
</tr>
</tbody>
</table>

This table summarizes the sample selection process for the HC disclosures used to train the *word2vec* model.
### TABLE 2 Seed Word List

<table>
<thead>
<tr>
<th>Category</th>
<th>Seed Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity, Equity, and Inclusion (DEI)</td>
<td>affirmative action, age, allyship, background, bias, black, color, DEI, disability, discriminate, discrimination, discriminatory, diverse, diversity, EEO, equal, equality, equitable, equity, ethnic, ethnicity, fair, female, gender, harassment, hispanic, human right, inclusion, inclusive, indigenous, injustice, justice, LGBT, LGBTQ, minority, nationality, opportunity, origin, pay gap, people of color, race, racial, religion, religious, representation, respect, respectful, sexual orientation, underrepresent, underrepresented, value, woman, women</td>
</tr>
<tr>
<td>Health and Safety</td>
<td>accident, COVID-19, death, fatality, hazard, health, healthy, illness, incident, injury, mental health, OSHA, pandemic, physical, protect, protection, respiratory, safe, safety, TRIR, vulnerable, well-being, wellness</td>
</tr>
<tr>
<td>Labor relations and culture</td>
<td>absenteeism, appraisal, attract, attraction, attrition, career, career development, child labor, childcare, coaching, code of conduct, collaborative, colleague, collective bargaining, commitment, core value, corrective, culture, develop, development, education, engagement, ethics, evaluate, feedback, forced labor, hire, hiring, labor disruption, mentor, mentoring, mentorship, motivate, organized labor, professional development, promotion, recruit, recruitment, relation, reskill, retain, retention, review, slavery, succeed, success, support, survey, talent, train, training, turnover, union, upskill, whistleblower, work environment, work stoppage</td>
</tr>
<tr>
<td>Compensation and Benefits</td>
<td>award, benefit, benefits, bereavement, bonus, compensation, contribution, equity award, health care, healthcare, hourly rate, incentive, insurance, maternity, medical, paid leave, parental leave, pay, pension, prescription drug, remuneration, restricted stock, retirement, salaries, salary, sick leave, stock option, vacation, wage</td>
</tr>
<tr>
<td>Demographics and others</td>
<td>associate, contract employee, demographic, demographics, employ, employee, full time, full-time, geographical region, geography, headcount, hourly, human capital, human resource, independent contractor, job function, management, part time, part-time, payroll, percent, personnel, region, salaried, sector, segment, skilled, staff, staffing, white-collar, worker, workforce</td>
</tr>
</tbody>
</table>
TABLE 3 HC Keyword List

<table>
<thead>
<tr>
<th>Category</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity, Equity, and Inclusion (DEI) (253 terms)</td>
<td>affinity_group, affirmative_action, affirmative_action_plan, african-american, african_american, age, alaska_native, alaskan_native, allyship, american_indian, ancestry, anti-discrimination, anti-harassment, anti-racism, anti-retaliation, asian-american, asian_american, background, belonging, bias, bias-free, bipoc, bisexual, black, brgs, bully, bullying, caucasian, cdo, corporate_equity_index, courtesy, creed, crgs, cross-cultural, cultural_background, dei, dei_council, deb, descent, dib, different_background, different_perspective, disability, disability_equity_index, disabled, discriminate, discrimination, discriminatory, discriminatory_harassment, disparity, diverse, diverse_background, diverse_candidate, diverse_candidate_slate, diverse_perspective, diverse_pool, diverse_slate, diversity, diversity-focused, diversity-related, domestic_partner, edi, eeo, eec, eigs, empathy, employment_opportunity, equal-opportunity, equal_employment, equal_employment_opportunity, equal_opportunity, equal_opportunity_employer, equal_pay, equal_treatment, equality, equality_index, equally, equitable, equitable_treatment, equitably, ergs, ethnic, ethnic_background, ethnic_diversity, ethnic_group, ethnic_minority, ethnic_origin, ethnically_diverse, ethnicity, fairness, female, gay, gei, gender, gender-equality, gender-neutral, gender_expression, gender_identification, gender_identity, gender_parity, generational, harassment, harassment_prevention, hawaiian, heritage, hispanic, hispanic-serving, historically_black_college, historically_underrepresented, human_right, human_right_campaign, ide, identity, implicit_bias, inclusion, inclusion-focused, inclusion-related, inclusive, inclusive_workplace, inclusively, inclusiveness, indigenous, individuality, inequality, inequity, injustice, intimidation, islander, justice, latino, latinx, lesbian, lgbt+, lgbtq, lgbtq+, lgbtq_equality, lgbtqa, listening_session, living_wage, male, marginalize, marital, marital_status, men, military_status, military_veteran, minority, minority_group, multi-cultural, multicultural, multigenerational, multiracial, naacp, national_origin, nationality, native_american, native_hawaiian, neurodiversity, non-caucasian, non-discrimination, non-discriminatory, non-minorities, non-minority, non-white, nondiscrimination, nondiscriminatory, pacific_islander, pay_equity, pay_gap, pay_parity, people_of_color, poc, political_affiliation, pregnancy, prejudice, protect_veteran, protect_veteran_status, protected_characteristic, protected_class, race, racial, racial_equality, racial_equity, racial_injustice, racial_justice, racial_minority, racially, racially_diverse, racism, reasonable_accommodation, religion, religious, religious_belief, representation, resource_group, respectful_conduct, respectful_workplace, respectfully, retaliation, retribution, self-disclosed, self-expression, self-identification, self-identified, self-identifies, self-identifying, sex, sexual, sexual_harassment, sexual_orientation, social_injustice, social_justice, socio-economic, socio-economic_status, socioeconomic, status_protect, stereotype, systemic_racism, traditionally_underrepresented, transgender, treat_fairly, unbiased, unconscious_bias, unconscious_bias_training, under-represented, under-represented_group, under-represented_minority, underprivileged, underrepresentation, underrepresented, underrepresented_community, underrepresented_ethnic, underrepresented_group, underrepresented_minority, underrepresented_population, underserved_community, unique_background, unique_perspective, unlawful_discrimination, urm, urms, values-based, varied_perspective, veteran, veteran-owned, veteran_status, veterans, without_bias, woman, women</td>
</tr>
<tr>
<td>Health and Safety (General) (157 terms)</td>
<td>acc, accident, accidental, accidental_death, aes, behavioral_health, biometric_screening, bls, case_rate, clean_supply, cleaning_procedure, cleaning_protocol, cleaning_supply, dart, death, depression, disinfect, disinfectant, disinfecting, disinfection, emergency_response, emotional, emotional_well-being, employee-hours, enhance_cleaning, enhanced_cleaning, ergonomic, ergonomics, fatality, first_aid, first_responder, fitness, fitness_center, fitness_class, flu_shot, flu_vaccination, frequent_cleaning, hand_sanitize_station, hand_sanitizer, handwashing, harm, hazard, hazardous, hazardous_material, health, health-related, health_authority, 56</td>
</tr>
</tbody>
</table>

Electronic copy available at: https://ssrn.com/abstract=4197489
<table>
<thead>
<tr>
<th>Category</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health and Safety (COVID) (70 terms)</td>
<td>cdc, cdc_guideline, coronavirus, coronavirus_disease, covid, covid-19, covid-related, daily_temperature_check, essential_worker, face-masks, face_cover, face_covering, face_mask, facemasks, international_non-essential_travel, mask, mask-wearing, mask_requirement, mask_wearing, masking, non-essential_travel, novel_coronavirus, pandemic, pandemic-related, physical_distance, physical_distancing, proper_social_distancing, quarantine, quarantined, quarantining, remote-work, remote-working, remote_work, remote_work_arrangement, remote_work_environment, remote_working, reopening, respirator, self-health, self-quarantine, self-screening, shelter-in-place, social-distancing, social_distance, social_distancing, social_distancing_guideline, stay-at-home, stay-at-home_order, telecommute, telecommuting, teleworking, temperature-taking, temperature_check, temperature_screen, temperature_screening, test_positive, travel_restriction, vaccinate, vaccination, vaccine, virus, virus_exposure, wear_face_covering, wear_mask, work-at-home, work_from-home, work_from-home_arrangement, work_from-home_guidance, work_from-home_policy, work_remotely, working_from-home</td>
</tr>
<tr>
<td>Labor relations and culture (362 terms)</td>
<td>absenteeism, abuse, abusive, advancement, afl cio, alpa, anonymous, anonymous_feedback, anonymous_hotline, anonymous_survey, anonymously, anti-bribery, anti-corruption, anticorruption, applicant, at-will, attendance_policy, attract, attracting, attraction, attracts, attrition, attrition_rate, bargain_unit, bargaining_group, bargaining_unit, best_employer, bribery, candid_feedback, candidate_pool, candidate_slate, candor, career, career-enhancing, career-oriented, career advancement, career development, career_growth, career_mobility, career_path, career_progression, charitable_cause, charitable_contribution, charitable_donation, charitable_foundation, charitable_gift, charitable_give, charitable_matching, charitable_organization, charity, child_labor, chro, co-worker, coaching, code_of_conduct, cohort, collaboration, collaborative, colleague, collective-bargaining_agreement, collective_bargaining, collective_bargaining_agreement, collective_bargaining_arrangement, collective_bargaining_unit, collegial, collegiality, commitment, community_involvement, community_outreach, commute, compassion, competency-based, compliance_hotline, confidential_hotline, confidential_reporting, confidentially, consultation, continual_learning, continuous_learn, continuous_learning, core_principle, core_value, corporate_philanthropy, corrupt, corruption, counseling, counseling_session, counselling, courage, courageous, coworkers, credential, cross-training, cultivate, cultural, cultural_awareness, cultural_fit, culture, customized_corporate_training, cwa, decency, dependent_care, direct_report, disaster_relief, eap, early-career, early_career, education, elearning, emerge_leader, emotional_intelligence, empathetic, employee-centric, employee-driven, employee-focused, employee-friendly, employee-led, employee-led_group,</td>
</tr>
<tr>
<td>Category</td>
<td>Keywords</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Compensation and Benefits</td>
<td>employee_value_proposition, employer_brand, employment-related, employment-related_work_stoppage, employment_decision, empowered, empowerment, engagement, engagement_score, engagement_survey, engaging, enjoyable_workplace, ethic, ethic_hotline, ethical, ethical_conduct, ethically, ethics, ethicsline, evaluation_process, executive-sponsored, exit_interview, external_hire, fair_treatment, fairly_compensate, family, fast-paced, flexible_schedule, forced_labor, fraud, fundraise, fundraising, furlough, furloughed, grievance, high-performance_culture, high_ethical_standard, highly_collaborative, hire, hiring, hiring_practice, honest, honestly, human_trafficking, humble, humility, iam, ibew, ibt, ilo, impasse, inappropriate_behavior, inspirational, inspiring, integrity, internal_candidate, internal_mobility, internal_promotion, international_brotherhood, international_union, internship, interview, interview_panel, interviewer, interviewing, involuntary_turnover, iuoe, job-related, job_specific, job_assignment, job_description, job_duty, job_fair, job_opening, job_satisfaction, labor-related, labor-related_work_stoppage, labor_disruption, labor_law, labor_relation, labour, laundering, lay-offs, layoff, leadership, learning, learning_management_system, like-minded, material_work_stoppage, mentor, mentoring, mentorship, mentorships, meritocracy, mid-career, moral, motivate, motto, new-hire, new_hire, newly_hire, non-profit_organization, non-unionized, nonprofit_organization, onboard, on-boarding, on-demand_learning, on-the-job, on-the-job_learning, on-the-job_training, onboarding, onboarding_process, one-on-one, online_learning, opieiu, open-door_policy, open_communication, open_dialogue, open_position, organized Labor, orientation, pafca, people-centric, people-first, performance_appraisal, performance_evaluation, personality, philosophy, post-employment, potential_successor, professional_development, professionalism, professionally, promote-from-within, promotion, pulse_survey, questionnaire, recruit, recruiter, recruiting, recruitment, reduction-in-force, reductions-in-force, renegotiation, representative_body, reprisal, resignation, reskill, reskilling, results-driven, retain, retain_qualified, retain_talented, retain_top, retaining, retains, retention, retention_rate, retentive, review_process, reward, rotational_assignment, self-assessment, self-development, self-directed_learning, self-paced_learning, significant_work_stoppage, skill-based, skill-building, skills-based, skillset, skillsets, slavery, soft_skill, solicit_feedback, succession, succession_plan, succession_planning, succession_planning_process, supervisor, survey, talent_acquisition, talent_attraction, talent_mobility, talent_pipeline, talent_pool, team-based, team-building, team-oriented, tenet, tenure, tenured, termination, top-performing, top_talent, town_hall_meeting, townhall, townhalls, trade_union, train, trained, trainee, trainer, training, turnover, turnover_rate, tutor, tutoring, uaw, umwa, unethical, unethical_behavior, union, union_contract, unionization, unionize, unionized, united_way, upskill, upskilling, uwua, values-based_culture, virtual_classroom, virtual_town_hall, voluntary_attrition, voluntary_attrition_rate, voluntary_separation, voluntary_turnover, voluntary_turnover_rate, volunteer, volunteer_activity, volunteer_hour, volunteering, vto, whistleblower, whistleblower_hotline, whistleblower_policy, work-life, work_life_balance, work_environment, work_stoppage, worker_council, working_environment, workplace</td>
</tr>
<tr>
<td>(283 terms)</td>
<td>401k, absence, accident_insurance, adoption_assistance, annual_bonus, annual_cash, annual_incentive, assistance_fund, assistance_program, autism, base_salary, base_wage, benefit, benefit_package, benefits, bereavement, bereavement_leave, bonus, bonus_award, bonus_payment, broad-based_equity, broad-based_equity_award, broad-based_stock_grant, broad-based_stock_incentive, care_provider, caregiver, caregiver_leave, caregiving, cash-based, cash-based_compensation_award, cash-based_performance_bonus, cash_bonus, cash_incentive, cash_incentive_plan, child-care, child_care, childcare, co-pays, commission, commission-based, commuter_benefit, company-matched, company-paid, company-paid_benefit, company-provided, company-subsidized, compensate_fairly, compensation, compensation_package, compensation_philosophy, competitive_base_salary, competitive_benefit_package, competitive_compensation, competitive_compensation_package, competitive_pay, competitive_salary, competitive_wage, complimentary, comprehensive_benefit_package,</td>
</tr>
</tbody>
</table>
### Category


<table>
<thead>
<tr>
<th>Demographics and others (160 terms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>administrator, advanced_degree, agent, apprentice, assistant_manager, associate, average_tenure, bachelor, back-office, clerk, co-op, college_graduate, college_student, consultant, contingent_worker, contract_employee, contract_worker, contractor, crew, crew_member, crewmember, crewmembers, demographic, demographics, district_manager, doctorate, electrical_worker, employ, employ_approximately, employed, employee, employee-related, engineer, entire_workforce, entry-level, entry_level, executive-level, executive_leadership, executive_officer, fixed-term, front-line, front-office, frontline, fte, full-time, full-time_basics, full-time_equivalent, full_time, full_time_equivalent, fulltime, general_manager, graduate_degree, headcount, high-caliber, high_caliber, high_performing, high_potential, highly-qualified, highly-skilled, highly_educate, highly_qualified, highly_qualify, highly_skill, highly_talented, highly_trained, hour_per_week, hourly,</td>
</tr>
</tbody>
</table>

Electronic copy available at: https://ssrn.com/abstract=4197489
<table>
<thead>
<tr>
<th>Category</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>hourly_team_member, hris, human-capital, human_capital, human_capital_management, human_resource, independent_agent, independent_consultant, independent_contractor, job_function, labor_cost, labor_force, labor_market, laborer, long-tenured, management-level, management_trainee, manager, manager-level, master_degree, mid-level_manager, motivated, name_executive_officer, non-employee, non-executive, non-management, non-represented, non-salaried, non-union, outside_consultant, part-time, part-time_basis, part_time, participation_rate, people, per_hour, per_week, personally, personnel, pre-employment, professional, qualified, qualified_applicant, qualified_individual, qualified_personnel, recent_graduate, salaried, sale_force, sale_representative, salesperson, senior-level, senior_executive, senior_leader, senior_leadership, senior_leadership_team, senior_management, senior_management_team, senior_manager, seniority, shift_schedule, skill, skilled, staff, staff_member, staffed, staffing, staffing_level, stagger_shift, stem-related, subcontractor, talent, talented_individual, team_member, teammate, teammates, technician, technicians, temporary_worker, unemployment, unskilled, unskilled_labor, vacancy, vocational, well-qualified, well-trained, work-related, work_schedule, workday, worker, workforce, workforce_demographic, workload, workweek</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 4 Definitions of Acronyms in Keyword List

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>American Chemistry Council</td>
</tr>
<tr>
<td>AES</td>
<td>Asbestos Environment and Safety</td>
</tr>
<tr>
<td>AFL-CIO</td>
<td>American Federation of Labor and Congress of Industrial Organizations</td>
</tr>
<tr>
<td>ALPA</td>
<td>Air Line Pilots Association</td>
</tr>
<tr>
<td>BIPOC</td>
<td>Black, Indigenous and People of Color</td>
</tr>
<tr>
<td>BLS</td>
<td>Bureau of Labor Statistics</td>
</tr>
<tr>
<td>BRGS</td>
<td>Best Employee Resource Groups</td>
</tr>
<tr>
<td>CDC</td>
<td>Centers for Disease Control and Prevention</td>
</tr>
<tr>
<td>CDO</td>
<td>Chief Diversity Officer</td>
</tr>
<tr>
<td>CHRO</td>
<td>Chief Human Resources Officer</td>
</tr>
<tr>
<td>CRGS</td>
<td>Colleague Resource Groups</td>
</tr>
<tr>
<td>CWA</td>
<td>Communications Workers of America</td>
</tr>
<tr>
<td>DART</td>
<td>Days Away Restricted or Transferred</td>
</tr>
<tr>
<td>DEI</td>
<td>Diversity, Equity and Inclusion</td>
</tr>
<tr>
<td>DEIB</td>
<td>Diversity, Equity, Inclusion, and Belonging</td>
</tr>
<tr>
<td>DIB</td>
<td>Diversity, Inclusion, and Belonging</td>
</tr>
<tr>
<td>EAP</td>
<td>Employee Assistance Program</td>
</tr>
<tr>
<td>EDI</td>
<td>Equality, Diversity and Inclusion</td>
</tr>
<tr>
<td>EEO</td>
<td>Equal Employment Opportunity</td>
</tr>
<tr>
<td>EEOC</td>
<td>Equal Employment Opportunity Commission</td>
</tr>
<tr>
<td>EIGS</td>
<td>Employee Inclusion Groups</td>
</tr>
<tr>
<td>ERGS</td>
<td>Employee Resource Groups</td>
</tr>
<tr>
<td>ESOP</td>
<td>Employee Stock Ownership Plan</td>
</tr>
<tr>
<td>ESPP</td>
<td>Employee Stock Purchase Plan</td>
</tr>
<tr>
<td>ETO</td>
<td>Earned Time Off</td>
</tr>
<tr>
<td>FTE</td>
<td>Full-time Equivalency</td>
</tr>
<tr>
<td>GEI</td>
<td>Gender-Equality Index</td>
</tr>
<tr>
<td>HRIS</td>
<td>Human Resources Information System</td>
</tr>
<tr>
<td>HSA</td>
<td>Health Savings Accounts</td>
</tr>
<tr>
<td>HSE</td>
<td>Health, Safety, and Environment</td>
</tr>
<tr>
<td>IAM</td>
<td>International Association of Machinists and Aerospace Workers</td>
</tr>
<tr>
<td>IBEW</td>
<td>International Brotherhood of Electrical Workers</td>
</tr>
<tr>
<td>IBT</td>
<td>International Brotherhood of Teamsters</td>
</tr>
<tr>
<td>IDE</td>
<td>Inclusion, Diversity &amp; Equity</td>
</tr>
<tr>
<td>ILO</td>
<td>International Labour Organization</td>
</tr>
<tr>
<td>IUOE</td>
<td>International Union of Operating Engineers</td>
</tr>
<tr>
<td>KSOP</td>
<td>A retirement plan that combines 401(K) and ESOP</td>
</tr>
<tr>
<td>LTCR</td>
<td>Lost Time Case Rate</td>
</tr>
</tbody>
</table>

Electronic copy available at: https://ssrn.com/abstract=4197489
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTI</td>
<td>Lost Time Injury</td>
</tr>
<tr>
<td>LTIP</td>
<td>Long Term Incentive Plan</td>
</tr>
<tr>
<td>LTIR</td>
<td>Lost Time Incident Rate</td>
</tr>
<tr>
<td>MSHA</td>
<td>Mine Safety and Health Administration</td>
</tr>
<tr>
<td>NAACP</td>
<td>National Association for the Advancement of Colored People</td>
</tr>
<tr>
<td>OHI</td>
<td>Other Health Impairment</td>
</tr>
<tr>
<td>OHSAS</td>
<td>Occupational Health and Safety Assessment Series</td>
</tr>
<tr>
<td>OPEIU</td>
<td>Office and Professional Employees International Union</td>
</tr>
<tr>
<td>OSHA</td>
<td>Occupational Safety and Health Administration</td>
</tr>
<tr>
<td>PAFCA</td>
<td>Professional Airline Flight Control Association</td>
</tr>
<tr>
<td>POC</td>
<td>People of Color</td>
</tr>
<tr>
<td>PPE</td>
<td>Personal Protective Equipment</td>
</tr>
<tr>
<td>PTO</td>
<td>Paid Time Off</td>
</tr>
<tr>
<td>QHSE</td>
<td>Quality, Health, Safety &amp; Environment</td>
</tr>
<tr>
<td>RIR</td>
<td>Reportable Incidence Rate</td>
</tr>
<tr>
<td>RSUS</td>
<td>Restricted Stock Units</td>
</tr>
<tr>
<td>SIF</td>
<td>Serious Injuries and Fatalities</td>
</tr>
<tr>
<td>TIR</td>
<td>Total Incident Rate</td>
</tr>
<tr>
<td>TRIR</td>
<td>Total Recordable Incident Rate</td>
</tr>
<tr>
<td>UAW</td>
<td>United Auto Workers</td>
</tr>
<tr>
<td>UMWA</td>
<td>United Mine Workers of America</td>
</tr>
<tr>
<td>URM</td>
<td>Underrepresented Minority</td>
</tr>
<tr>
<td>URMS</td>
<td>Underrepresented Minorities</td>
</tr>
<tr>
<td>UWUA</td>
<td>Utility Workers Union of America</td>
</tr>
<tr>
<td>VPP</td>
<td>Voluntary Protection Programs</td>
</tr>
<tr>
<td>VTO</td>
<td>Volunteer Time Off</td>
</tr>
</tbody>
</table>