JOURNAL OF INFORMATION SYSTEMS
Vol. XX, No. XX
MONTH YEAR
pp. 1–13

American Accounting Association
DOI: 10.2308/ISYS-2023-003

Employee Ratings and Reviews Data from Glassdoor

Mi (Jamie) Zhou
Virginia Commonwealth University

Yaxuan (Ashley) Li
Idaho State University

Zhilei (George) Qiao
The University of Alabama at Birmingham

Bowen Shi
Central University of Finance and Economics

ABSTRACT: This paper presents the employee ratings and reviews data from Glassdoor and the R codes used to collect, clean, and organize the data. We collect three types of information for each Glassdoor review: review metrics, content, and reviewer information. We also calculate some commonly used textual metrics, such as sentiment, readability, the number of uncertainty words, etc. The datasets include necessary identifiers that can connect to other financial data sources. All the variables and metrics are provided at the review level, which enables researchers to aggregate the data from different levels and angles. As a demonstrative example, we use a simple word list to measure how often employees mention COVID-19 in review comments. The R codes provided as an RStudio project are self-contained and can also be modified and applied to other data sources of interest.

Keywords: employee ratings; employee reviews; Glassdoor; R; RStudio; textual metrics.

I. INTRODUCTION

Employees possess valuable information about their employers, which makes understanding employees' perceptions of their firms an important task. Their unique information can be utilized not only by management in budgeting and planning, but also by investors and analysts in external disclosure (Huang, Li, Meschke, and Guthrie 2015). Consequently, we have witnessed scholars across many business disciplines seeking to understand how employees perceive their jobs in different settings. Research in management, marketing, information system, finance, and economics suggests that employees' perceptions provide insights into firm operation strategy and firm performance (Cooper, Diab, and Beeson 2020; Green, Huang, Wen, and Zhou 2019; Karabarbounis and Pinto 2018; Saini and Jawahar 2019). Scholars of accounting are increasingly exploring employee information regarding its impact on firm performance, financial reporting risks, audit outcomes, corporate information environment, and corporate financing and investment policies (Chemmanur, Rajaiya, and Sheng 2020; Hales, Moon, and Swenson 2018; Huang, Li, and Markov 2020; Huang, Masli, Meschke, and Guthrie 2017; Lei, Li, and Luo 2019).

We thank Michael Werner (editor) and anonymous reviewers for their insightful comments and constructive suggestions. The authors of this publication have no conflicts of interest related to this research.

Mi (Jamie) Zhou, Virginia Commonwealth University, School of Business, Department of Accounting, Richmond, VA, USA; Yaxuan (Ashley) Li, Idaho State University, College of Business, Department of Accounting and Information Systems, Pocatello, ID, USA; Zhilei (George) Qiao, The University of Alabama at Birmingham, Collat School of Business, Department of Management, Information Systems and Quantitative Methods, Birmingham, AL, USA; Bowen Shi, Central University of Finance and Economics, School of Innovation and Development, China Economics and Management Academy, Beijing, China.

Supplemental materials are available online, as linked in the text.

Editor's note: Accepted by Michael Werner, under the Senior Editorship of Tawei (David) Wang.

Submitted: January 2023
Accepted: June 2024
Early Access: August 2024

Research on employee information traditionally focuses on top executives and managers, usually via interviews, surveys, or analysis of insider trading information (Li, Minnis, Nagar, and Rajan 2014; Huang, Teoh, and Zhang 2014; Allee and Deangelis 2015). This approach is both intuitive and practical because upper-level management has largely controlled firm-level communications, such as press releases, earnings warnings, and other such announcements (Hales et al. 2018). Although employees across an organization often possess extensive knowledge about their company's condition, studies on such information are limited (Huddart and Lang 2003; Babenko and Sen 2016). One of the primary challenges in exploring rank-and-file employee information is the gathering of information, as historically, there have been limited avenues for sharing this information outside the organization (Hales et al. 2018). Sites and applications focused on investing and firm disclosures do not typically collect or provide employment information. As a result, data limitations have restricted empirical researchers' ability to analyze rank-and-file employee information.

To tackle this challenge and facilitate understanding of employee information, this paper presents the ratings and reviews data from Glassdoor, along with the R codes used to collect, clean, and organize the data. Glassdoor opinions are offered by rank-and-file employees rather than from the general public and therefore are potentially a rich source for private or qualitative information about the working condition for employees or of employee mood (Teoh 2018). For each Glassdoor review, we collect three types of information: review metrics, review content, and reviewer information. Review metrics are different employee ratings, including an overall rating of the firm, as well as ratings on five specific areas: culture and values, work/life balance, senior management, compensation and benefits, and career opportunities. Review content includes textual comments (pros, cons, and advice to management) made by employees about the employer or the job, the review date, and the number of people who found this review helpful. Reviewer information includes the characteristics such as employee job title, tenure, location, etc. We also calculate some commonly used textual metrics, such as sentiment, readability, and the number of uncertainty words. The dataset is provided with necessary identifiers (GVKEY, TICKER, CUSIP, etc.) that can be used to connect other financial data sources, such as Compustat, CRSP, etc. All the variables and metrics are provided at the review level, which enables researchers to aggregate and study the data from different levels and angles. The R codes are provided as a self-contained RStudio project that can be executed "out of the box"; the codes can also be modified and applied to other data sources of interest.

The paper is organized as follows. In Section II, we provide a brief background of research on employee disclosure, including what Glassdoor is and how it is used in several recent research. In Section III, we introduce the structure of the dataset and discuss the steps used to collect, clean, and organize the data. In Section IV, we present the R codes used to generate the dataset along with packages used to complete the necessary steps. A discussion of potential research opportunities is presented in Section V, where we demonstrate how our code and dataset might be used by accounting researchers in different settings. The paper concludes in Section VI with a summary of the dataset and methodology.

II. BACKGROUND

In the last few years, online social media has become an important channel for distributing and gathering corporate information. Due to the rise of social media websites focused on job searching, employees can now communicate and share their information with others. One of the popular websites containing reviews and ratings generated by employees is Glassdoor.

Glassdoor is a widely used recruiting and social media site for employees and job seekers. It attracts 67 million unique visitors per month to its 12 million job listings as well as its 114 million company reviews and compensation insights (Dube and Zhu 2021; Glassdoor 2022). Glassdoor.com was founded by Robert Hohman, Rich Barton, and Tim Besse in 2007 (Glassdoor 2022). It is a website where employees and former employees voluntarily and anonymously review their companies. To use the service, registered users are asked, among other things, to report their current occupation title (job position), company, salary (in addition to other payment schemes), location, and level of experience. In return, users can get access to user-generated content, including employee ratings and reviews of companies, interview questions, CEO approval rates, and summary statistics of salaries for job positions within each company (Glassdoor 2022, 2024).

Glassdoor reviews are organized by an employer (company)-page-review structure. One employer can have many pages of reviews, and one page contains ten reviews. For each individual review, Glassdoor provides several metrics to gauge employee opinion. Most salient among the metrics is the Overall Rating, which ranges from 1 to 5. Each employee is also given the option to rate their employer along with six additional aspects: Work/Life Balance, Culture and Values, Diversity & Inclusion, Career Opportunities, Compensation and Benefits, and Senior Management. Some general employee information (employee status and tenure) is displayed under the ratings. Each review has a title, followed by the review date and employee information, such as job title and location. There is an open-form section of the review that allows the reviewer to state in his/her own words any other comments about the employer or the job that he/she would like to include. For each review, Glassdoor also asks the employee to provide ratings (three scales) on (1)

Recommendation to a Friend (yes, neutral, and no), (2) CEO Approval (approval, neutral, and disapproval), and (3) Business Outlook (positive, neutral, and negative). Figure 1 shows those six ratings. In the textual comments, reviewers (employees) are suggested to comment on three aspects: Pros, Cons, and Advice to Management. Lastly, similar to the "like" feature used in other social media platforms, Glassdoor also records the number of people who "found this review helpful." Please note that not all the employees provide ratings for all the aspects or write their reviews in the suggested format by Glassdoor. A typical employee review is illustrated in Figure 1.

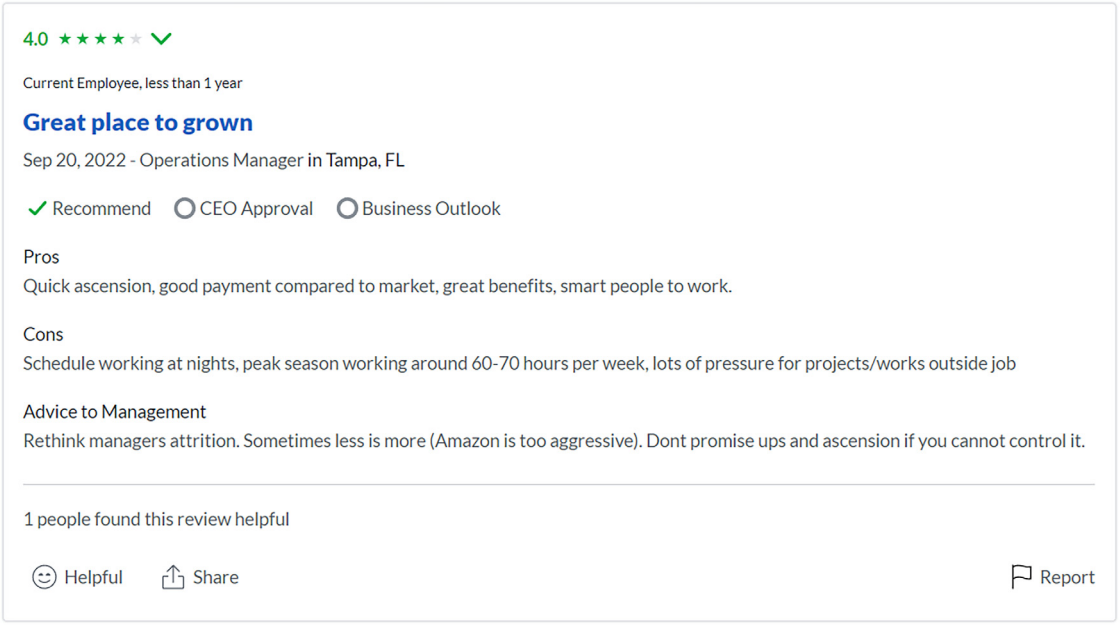
Glassdoor has been used in many recent studies on employee disclosures. Researchers use Glassdoor CEO approval rating as a proxy for the observability of CEOs' self-serving activities to study how subordinates can discipline CEOs' such activities (Li 2019). Specifically, the author assumes that firms where more employees "approve" or "disapprove" of their CEOs have higher observability than firms where more employees have "no opinion of their CEOs." Using Glassdoor ratings on senior management and overall firm, researchers also examine whether tax avoidance news negatively affects employee perceptions of managers and firms (Lee, Ng, Shevlin, and Venkat 2021). The authors examine the textual content of reviews to determine whether tax news increases the discussion of taxes by counting the number of Glassdoor reviews mentioning "tax" or "taxes" in the Cons and Advice to Management sections of the reviews. Last, but not least, researchers conduct analysis for the determinants of employer ratings and find that employer rating is significantly positively related to size, turnover, and returns over the previous six months and negatively related to book-to-market, institutional ownership, and insider trading (Green et al. 2019). The authors also find that the "100 Best Companies to Work For" status is significantly related to Glassdoor ratings.

III. DATA COLLECTION APPROACH AND DATASET STRUCTURE[1]

Data Collection

The data collection approach is straightforward. First, for each of our sample companies, we manually search for its Glassdoor profile and then its landing page of Glassdoor reviews (i.e., the home page of a firm's employee reviews).

FIGURE 1
An Example of Employee Reviews at Glassdoor.com



(The full-color version is available online.)

[1] For additional information, see the supplemental material. The completed Glassdoor review datasets and the relevant R codes can be found in the Datasets_Codes.zip folder, and the description of codes and datasets can be found in the file CodesAndDatasetInformation.docx.

The URLs of Glassdoor profiles (and landing page of reviews) consist of company names and numbers (possibly internal identifiers created by Glassdoor). We do not find a way to automate this step, as there are variations of a company's name, and we do not know what "internal" number is assigned to the given company. Second, for each landing page of reviews, we use the default setting to limit reviews by language (English only) and type (full time and part time). Third, we scrape (download) all review pages for each company and save them on a local computer drive. In other words, we navigate and send web requests for each review page, then save them locally as individual HTML files. Glassdoor imposes a rate limit to block high-frequency access to its web pages. We send our request from multiple computers with different internet protocol (IP) addresses and intentionally stop scraping after several hours to bypass the rate limit. In late 2023,[2] Glassdoor imposed stricter antiscraping mechanisms that detect and block direct data scraping through automatic bot scripts. More advanced methods, such as Selenium browser automation tools, IP rotation, and web scraping services (Scraping Robot 2022; Jayan 2023; Bright Data 2024) that can mimic human browsing activities, should be used to access and download review pages. Fourth, we investigate the page content and parse the required information from review pages. In this step, we loop through each saved HTML file and extract all required review data, such as employee information, review content, and so on. Fifth, we clean, split, standardize, and organize the data. As not all the information is supplied by employees, necessary cleaning procedures are used to deal with empty values, nonrecognized characters, etc. In addition, some information is extracted and separated into more reasonable and readable data variables/fields. For example, as shown in Figure 1, employee type and job tenure are provided in one sentence "Current Employee, less than 1 year"; necessary procedures are used to separate such information and store them into respective data fields. Similarly, the review date, job title, and location are extracted and put in different data fields.

Furthermore, visual indicators/icons are converted into corresponding numerical values. For example, employee ratings of Recommend, CEO Approval, and Business Outlook are shown by using check marks, which are converted to 1, 0, and −1, respectively. We also use the dictionary and formula used in previous research, especially in accounting and finance areas (Loughran and Mcdonald 2014, 2016), to calculate some common textual metrics, such as sentiment, readability, and the number of uncertainty words. Finally, we review the dataset and run some descriptive analysis to check and address potential errors and anomalies.

We organize the dataset according to the structure provided by Glassdoor, with necessary modifications for better use of data. Our final dataset contains both the data from Glassdoor reviews and some derived variables. They can be roughly classified into six types, including identifier information, employee information, employee ratings, review content, sentiment, and other textual metrics. The variables and metrics provided at the review level make it easy for researchers to aggregate and study the data from different levels and angles. For example, researchers can focus on part-time employees or employee opinions on a specific topic during a certain period. Our data variables and metrics are summarized in Table 1.

Data Description

Our dataset covers employee reviews as of September 30, 2021 for S&P 1500 companies. We started in January 2021 by searching for the companies' Glassdoor profile pages and landing pages of employee reviews. It is possible that some companies did not have a profile page or employee reviews on Glassdoor when the list of profile pages was manually searched and compiled. For example, the company WageWorks (GVKEY 187105) had its first employee review on April 1, 2022. As a result, reviews of the company WageWorks were not included in our dataset.

In order to reduce potential errors and maintain consistency, we double check the landing pages of employee reviews to remove duplicated, invalid, and incorrect landing pages. In the case of mergers and acquisitions, companies may keep their reviews separate or combine their reviews together. We keep the review data separated by original company profile (via GVKEY) and let the data users decide whether and when to combine review data for those companies. In addition, it is possible that one company has multiple profiles (thus multiple landing pages of reviews) created on Glassdoor. One of the most common reasons is that occasionally employer users inadvertently create multiple Glassdoor accounts using various email addresses (Glassdoor 2024). In this case, we use our judgment and choose a profile as the primary one based on the number of employee reviews and the date of most recent reviews. Finally, we spot check the top and bottom 1 percent of companies to make sure their employee reviews are correctly scraped.

Our final sample selection and review data are described in Table 2. Panel A presents the sample selection process. We started with the S&P 1500 company list from Compustat. There are 26 companies for which we could not find a Glassdoor profile, and there are 12 companies with no employee reviews posted. Our final sample contains 1,462 companies. Panel B presents the company distribution by number of employee reviews. There are 1,780,538 total employee

[2] We found this in December 2023 by noticing that our previous scraping codes did not work anymore.

TABLE 1

Summary of Variables and Metrics

Variable Type	Variable/Metric Description
Review Identifier	Company GVKEY
	Company name
	Review ID
	Review date
Employee Information	Employee type (current, former, full-time, part-time, etc.)
	Employee tenure (number of years)
	Employee title
	Employee location
Employee Ratings (5-point scales)	Overall Rating
	Work/Life Balance
	Culture & Values
	Diversity & Inclusion
	Career Opportunities
	Compensation and Benefits
	Senior Management
Review Content	Recommend
	CEO Approval
	Business Outlook
	Pros (textual narrative)
	Cons (textual narrative)
	Advice to Management (textual narrative)
	Number of people found this review helpful (numerical value)
Sentiment (Tone)	Pure positive sentiment
	Pure negative sentiment
	Net sentiment
Other Textual Metrics	Number of words
	Number of sentences
	Number of negative words
	Number of positive words
	Number of uncertainty words
	Number of litigious words
	Number of strong modal words
	Number of weak modal words
	Number of constraining words

reviews from 1,462 companies. Specifically, there are 29 (2.0 percent) companies that have 10,000 or more employee reviews, together having 682,626 (38.3 percent) employee reviews, and 75 (5.1 percent) companies that have 5,000 or more employee reviews, together having 996,927 (56.0 percent) employee reviews. As shown in Panel B, the number of employee reviews vary greatly by company. The top 75 companies have more than half (56.0 percent) of the total employee reviews, and the last 439 (1,462 subtracted by 1,023) companies have just a little over half a percent (0.6 percent, 100 percent subtracted by 99.4 percent) of the total reviews.

Table 2, Panel C presents the numbers of companies and employee reviews by industry based on two-digit SIC code. As expected, the numbers of companies and employee reviews vary greatly by industry. Specifically, the manufacturing industry has the most companies (577, 39 percent of the total 1,462), with 364,130 (20.5 percent of total 1,780,538) employee reviews. However, the retail trade industry has the most employee reviews (498,984, 28.0 percent), from 105 (7.2 percent) companies.

Table 2, Panel D presents the numbers of companies and employee reviews by year from 2008 to 2021 (as of September 30, 2021). As shown, the number of companies with at least one review and the number of employee reviews

TABLE 2

Sample Section and Description

Panel A: Company Selection

Description	Number
The S&P 1500 company list from Compustat	1,500
Less companies with no profile found	26
Less companies with no employee reviews posted	12
Total	**1,462**

All numbers above are based on steps completed by January 2021.

Panel B: Distribution by Number of Employee Reviews

Description	Number of Companies	Accumulative Percent of Companies	Number of Reviews	Accumulative Percent of Reviews
At least 10,000 employee reviews	29	2.0	682,626	38.3
At least 5,000 employee reviews	75	5.1	996,927	56.0
At least 1,000 employee reviews	324	22.2	1,558,433	87.5
At least 500 employee reviews	467	31.9	1,661,944	93.3
At least 100 employee reviews	825	56.4	1,755,578	98.6
At least 50 employee reviews	1,023	70.0	1,769,284	99.4
At least 1 employee review	1,462	100.0	1,780,538	100.0

All numbers above are based on employee reviews collected by September 30, 2021.

Panel C: Number of Companies and Employee Reviews by Industry (Based on Two-Digit SIC Code)

Year	Number of Companies	Percent of Companies	Number of Reviews	Percent of Reviews
A. Agriculture, Forestry, & Fishing	2	0.1	197	0.0
B. Mining	62	4.2	19,132	1.1
C. Construction	25	1.7	4,779	0.3
D. Manufacturing	577	39.5	364,130	20.5
E. Transportation & Public Utilities	124	8.5	154,234	8.7
F. Wholesale Trade	45	3.1	22,562	1.3
G. Retail Trade	105	7.2	498,984	28.0
H. Finance, Insurance, & Real Estate	319	21.8	289,414	16.3
I. Services	200	13.7	406,852	22.8
K. Nonclassifiable Establishments	3	0.2	20,254	1.1

All numbers above are based on employee reviews collected by September 30, 2021.

(continued on next page)

both increase over time. There is a noticeable big jump in the number of employee reviews from 2013 to 2014, then the number fluctuates up and down in recent years.

IV. R CODES

In this section, we present our R codes for the data collection approach described in the previous section, along with an example of developing a COVID-19-related measure using a customized word list (e.g., COVID-19, pandemic, coronavirus, mask, quarantine, and social distancing). The R codes are provided as a complete R project created using RStudio and can be downloaded to any computer. Table 3 illustrates the tree structure of the project folder, assuming the project files are stored at "D:\JIS_GLASSDOOR_DATA." Panel A shows the initial tree structure of the project folder, and Panel B shows the tree structure after a demo run is completed.

TABLE 2 (continued)

Panel D: Number of Companies and Employee Reviews by Year

Year	Number of Companies	Percent of Companies	Number of Reviews	Accumulative Percent of Reviews
2008	545	37.3	9,271	0.5
2009	626	42.8	14,811	1.4
2010	727	49.7	28,326	2.9
2011	761	52.1	35,292	4.9
2012	818	56.0	58,169	8.2
2013	831	56.8	77,790	12.6
2014	880	60.2	128,469	19.8
2015	931	63.7	214,732	31.8
2016	932	63.7	231,829	44.9
2017	949	64.9	228,442	57.7
2018	980	67.0	196,199	68.7
2019	1,042	71.3	195,672	79.7
2020	1,264	86.5	260,303	94.3
2021	1,365	93.4	101,233	100.0

All numbers above are based on employee reviews collected by September 30, 2021.

TABLE 3

Tree Structure of Project Folder

Panel A: Initial Structure

```
D:\JIS_GLASSDOOR_DATA
|   JIS_Glassdoor_Data.Rproj
|   JIS_Glassdoor_Data_R_Codes.Rmd
|   JIS_Glassdoor_Data_R_Codes_LocalDemo.Rmd
|
+---input
        JIS_Glassdoor_landingPages.csv
        LM_Dic_1993_2021.rds
        textualVars.R
|
+---output_demo
    \---JIS_reviewpages
            reviews_14898_3D-Systems_1.html
            reviews_14898_3D-Systems_2.html
            reviews_14898_3D-Systems_3.html
            reviews_65406_1800PetMeds_1.html
            reviews_65406_1800PetMeds_2.html
```

Panel B: Structure after Demo Run

```
D:\JIS_GLASSDOOR_DATA
|   JIS_Glassdoor_Data.Rproj
|   JIS_Glassdoor_Data_R_Codes.Rmd
|   JIS_Glassdoor_Data_R_Codes_LocalDemo.Rmd
|   JIS_Glassdoor_Data_R_Codes_LocalDemo.html
|
+---input
|       JIS_Glassdoor_landingPages.csv
|       LM_Dic_1993_2021.rds
|       textualVars.R
|
\---output_demo
    |   JIS_Glassdoor_individualpage_urls.csv
    |   JIS_Glassdoor_reviews_data_raw.csv
    |   JIS_Glassdoor_review_data_clean.csv
    |   JIS_Glassdoor_review_data_clean.rds
    |   JIS_Glassdoor_review_data_covid.csv
    |   JIS_Glassdoor_review_data_covid.rds
    |   JIS_Glassdoor_review_data_text_metrics.csv
    |   JIS_Glassdoor_review_data_text_metrics.rds
    |
    \---JIS_reviewpages
            reviews_14898_3D-Systems_1.html
            reviews_14898_3D-Systems_2.html
            reviews_14898_3D-Systems_3.html
            reviews_65406_1800PetMeds_1.html
            reviews_65406_1800PetMeds_2.html
```

The file JIS_Glassdoor_Data.Rproj is the project file created by RStudio. Required initial data (e.g., the Glassdoor landing pages of each firm) and resources (e.g., sentiment dictionaries and utility functions) are included in the folder named "input." The R codes used to scrape, clean, and organize review data are contained in an R Markdown file (JIS_Glassdoor_Data_R_Codes.Rmd). The R Markdown file is organized into different parts with detailed comments. Each part of the codes will be discussed in more detail in the following subsection.

Once the R project is downloaded to a local computer, it can be loaded into the RStudio development environment by double clicking the project file. Because of Glassdoor's antiscraping mechanism, now we have to use the Selenium method to scrape data. In order to use the Selenium method, users must have JAVA and Firefox browser installed in their computer (in addition to R, RStudio, and other packages), and some user interactions (e.g., logging in Glassdoor website) are also required to run the codes. As a result, the R program file (JIS_Glassdoor_Data_R_Codes.Rmd) cannot be run directly and needs to executed chunk by chunk with necessary user interactions in Part 1. When the codes are executed, review pages will be scraped and parsed review data will be placed in a folder named "output," which, if it does not exist, will be automatically created by the R codes. To avoid intensive data scraping from Glassdoor.com, the initial codes are preset to scrape and parse only the first five review pages (roughly 50 employee reviews). Our codes can also be modified and applied to other data sources of interest.

We included another R program file (JIS_Glassdoor_Data_R_Codes_LocalDemo.Rmd) for demonstration purposes. The demo R program assumes Part 1 is completed, and some review pages are downloaded to a local folder named "output_demo." The demo R program can be executed directly using the "Knit to Html" feature in RStudio. When the demo R Markdown file is executed, an HTML file (JIS_Glassdoor_Data_R_Codes_LocalDemo.html) is generated. The HTML file is a code documentation of the R Markdown file; it contains structured comments, explanations, and the sample output data for important steps of data collection, cleaning, and metric calculations.

Part 1: Prepare Review Page URLs and Scrape Review Pages to Local HTML Files

The file JIS_Glassdoor_landingPages.csv contains the landing page links of the S&P 1500 firms. This file is manually prepared by searching each firm's name within Glassdoor.com; we also record the total number of reviews, which can be found on each landing page. Employee reviews at Glassdoor.com are organized by page, and each page has ten reviews (except the last page, which may have less than ten reviews). After knowing the total number of reviews for a given firm, we can construct all the individual page links by adding the page information to the link. For example, the landing page for firm AAR Corp. is https://www.glassdoor.com/Reviews/AAR-Reviews-E4.htm; when page information is added, the link to its second page of employee reviews is https://www.glassdoor.com/Reviews/AAR-Reviews-E4_P2.htm. From the landing page, we can find 468 reviews (this number changes when new reviews are added). We need to construct 46 individual review pages. Along with the landing page, we have a total of 47 review pages to scrape.

After the individual page links are prepared, we use R codes to scrape review pages as HTML files and save them into a subfolder ("output\JIS_reviewpages") inside the project folder. If an employee review is very long, Glassdoor only displays it partially. The rest of the review can be dynamically loaded by clicking the text/button "shows more" or "continue reading." Since user interaction is needed to load the complete review data, we use the Selenium package to automatically mimic the user interaction with the review pages. Once all the review data are loaded, the review page is saved into the subfolder mentioned above.

Part 2: Parse the Review Pages Scraped in HTML Format

We use the *rvest* package to parse out information from each review item. As illustrated in Figure 1, each review has three types of information: review metrics, review content, and reviewer information. Review metrics are different employee ratings, including an overall rating of the firm, as well as ratings on five specific areas: culture and values, work/life balance, senior management, compensation and benefits, and career opportunities. Review content includes textual comments (pros, cons, and advice to management) made by employees about the employer or the job, as well as review date and the number of people who found this review helpful. Reviewer information includes characteristics such as employee job title, tenure, location, etc.

These three types of information are not clearly separated in underlying HTML files (e.g., not in separate HTML tags). It is normal to see several pieces of information combined and presented together. For example, the short sentence "Current Employee, less than 1 year" shown in Figure 1 contains employee status and tenure. In addition, the review date and job title are also presented together under the review title. In this step, we parse the HTML files and extract those review data "as it is" and save it into a file named JIS_Glassdoor_reviews_data_raw.csv. This raw data file is then used as the starting point to clean and organize the review data.

Based on the raw data file, we separate information pieces presented together, assign each information piece a meaningful variable name, and convert them into the correct data type. For example, the date information and job tenure should be further extracted from those textual sentences and converted to date and number, respectively. The value of "Recommend," "CEO Approval," and "Business Outlook" are presented as visual check mark icons. Similarly, the rating on each of the five specific areas is shown as the number of stars in black color. Those icons and stars should be extracted and converted into correct values. In addition, new variables could be derived from the review data. For example, employee type (current employee, former employee, intern, etc.) can be derived to represent the reviewer's current relationship with the firm. The full list of variables is provided in Table 1. The clean and organized review data are provided in the file JIS_Glassdoor_review_data_clean.csv. The clean data are also saved as an R data file (in rds format) to preserve the data type information better.

Part 3: Calculate Additional Textual Metrics

Although there are some ratings in the review data, most of the information is contained in the review text. Lots of textual analyses could be applied to the review text; one of them is to extract features from text using the "Bag-of-Words" approach. A bag-of-words is a representation of text that describes the occurrence of features words (e.g., sentiment words) within a document (e.g., review text). It involves two things: a vocabulary of known words (e.g., happy and sad) and a measure of the presence of known words (e.g., frequency). It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The approach is only concerned with whether known words occur in the document, not where in the document (Brownlee 2019). We use the *SentimentAnalysis* package to extract and construct different sentiment variables based on a variety of popular sentiment dictionaries (i.e., Harvard IV and Loughran-McDonald). The *SentimentAnalysis* package can also be used to extract textual features other than sentiment if a customized dictionary is provided. Our codes use several popular Loughran-McDonald dictionaries to construct textual metrics regarding uncertainty, litigation, constrain, strong modal, and weak modal. We provide the number of dictionary words found in each review along with the total number of words and sentences. This gives users the flexibility to calculate the textual metrics in different ways. As a result, alternative metrics based on different dictionaries and different formulas (e.g., scaled by all the words or only the sentiment words) can be added when deemed necessary. The textual metrics are provided in the file JIS_Glassdoor_review_data_clean.csv; they are also saved as an R data file (in *rds* format) to better preserve the data type information.

Part 4: Demonstrative Codes to Generate COVID-19-Related Reviews

As an extension to textual metrics based on widely used dictionaries, we define a simplified but customized COVID-19 dictionary and construct a metric indicating whether the review is related to COVID-19. Our COVID-19 dictionary contains words such as COVID-19, pandemic, coronavirus, mask, quarantine, social distancing, and so on. Our codes use word/phrase matches using regular expression; if an employee review contains any of those words, the review is COVID-19 related. The COVID-19 indicators are provided in the file JIS_Glassdoor_review_data_COVID-19.csv. Similarly, they are also saved as an R data file (in rds format) to better preserve the data type information.

The COVID-19 feature can also be based on different parts of employee review text. For example, in addition to searching COVID-19-related words from all review text, we can only search from the "Pros," "Cons," or "Advice to Management." In this way, the COVID-19 indicator is related to a specific part of the review. If a COVID-19 keyword is found in the Pros or Cons part of the review, it is likely the reviewer has a positive or negative feedback about the COVID-19 practice of the firm, respectively. Here, the key is to construct a reasonable dictionary representing COVID-19-related topics. Once the COVID-19 keywords are identified, we can construct a customized dictionary. Just like the popular Loughran-McDonald dictionaries, this newly constructed COVID-19 dictionary can be used in the *SentimentAnalysis* package mentioned above to generate COVID-19-related reviews. Similarly, we can extract textual features related to other topics as long as a relevant dictionary can be constructed.

V. POSSIBLE FUTURE DIRECTIONS

Glassdoor provides rank-and-file employees a platform to voluntarily and anonymously express opinions about their employers (companies); such opinions could later be conveyed in their employers' voluntary disclosures and mandated accounting reports (Hales et al. 2018). As a result, accounting researchers can use Glassdoor employee reviews as a new information channel and conduct interesting studies related to both voluntary and mandatory disclosures. Current research using Glassdoor review data is mainly using numerical metrics; the textual content of reviews is less studied. A future direction for the research using Glassdoor data may be to conduct textual analysis of employee

opinions to obtain deeper insights about the drivers of employee and team productivity (Teoh 2018). Our Glassdoor dataset provides a wealth of information about employee feedback and perceptions and could provide a range of research opportunities for accounting researchers. As this dataset can be used in a variety of accounting settings, there are several possibilities for future research.

The dataset presented in this paper could be beneficial for accounting researchers conducting studies on the role of accounting information in labor markets. deHaan, Li, and Zhou (2023) investigate whether firms' public financial reports (earnings announcements) cause their rank-and-file employees to reevaluate their jobs and consider leaving. They use weekly counts of Glassdoor reviews by current employees as a proxy for new job searchers and use abnormal within-quarter changes in review counts to identify changes in search around earnings announcements. They find that job searches by current employees increase significantly during earnings announcement weeks, especially when employees are more mobile and when their information frictions are greater. By using the textual content of employee reviews, researchers can investigate the underlying reasons why those employees are leaving. Researchers have also used Glassdoor reviews to study the implications of employee satisfaction and work-life balance (Hammami, Moldovan, and Peltier 2020; Khavis and Krishnan 2021; Khavis, Krishnan, and Tipton 2022; Hope, Li, Lin, and Rabier 2021). For example, Hammami et al. (2020) examine how an auditor's salary perception impacts their audit quality and delay; Khavis and Krishnan (2021) show that firms' internal characteristics can explain employee satisfaction and audit quality. These studies can also benefit from the text descriptions of employee reviews as such internal characteristics may have been mentioned by employees.

It is also possible for accounting researchers to analyze employee feedback concerning major events that are occurring within their organizations, including tax policy changes, upcoming investment opportunities, changes in executive teams, significant litigation proceedings, adoption of emerging technologies, and so on. Using employee satisfaction data from current and former rank-and-file employees from a source similar to Glassdoor, Shan and Tang (2023) find that companies with higher employee satisfaction scores withstand COVID-19 better. In the above section, we demonstrate how to generate COVID-19-related reviews using a simplified but customized dictionary as well as a metric. Researchers who are interested in COVID-19-related comments could gain a detailed understanding of employee feedback in this particular area, as well as conduct pre- and postevent comparisons and other additional studies. Using our code, accounting information system researchers could locate and track the impact of other major events, including security breaches, blockchain adoption, XBRL implementation, automated audits, etc.

As an additional application, the Glassdoor dataset can be used as a proxy for quantifying organizational culture. By leveraging the Glassdoor dataset, researchers could identify the strengths and weaknesses of a company's culture and workplace dynamics, as perceived by employees. This information could then be used to develop interventions to improve the company's culture, which could lead to increased employee satisfaction, productivity, and overall organizational performance. Employees are important stakeholders, and researchers can use the Glassdoor dataset to examine the relationship between employee perceptions and other variables of interest, such as customer satisfaction and brand engagement. By analyzing these relationships, researchers could gain a better understanding of the factors that contribute to employee satisfaction and how these factors impact other aspects of organizational performance. For example, Glassdoor ratings can provide insights into how employer branding impacts employee perceptions of their employers. And researchers can examine how a company's culture and values impact employee engagement and productivity and how these, in turn, impact the company's ability to attract and retain talent.

As a noncorporate disseminated data source, Glassdoor data can be used to supplement and validate information generated by companies. Hales et al. (2018) find that Glassdoor employee outlook predicts financial statement line items (e.g., sales, gross margin, and earnings) as well as earnings surprises and management guidance. One direction for possible extension to their research is to examine whether the predictability of the outlook variable is incremental to other sources or channels of information. Once again, the textual content of employee reviews can be studied to provide insight about whether reviews produce otherwise unavailable information or merely reflect other information that may already be publicly available, such as from past history of earnings surprise and analyst revisions (Teoh 2018).

Researchers also use Glassdoor employee ratings of senior managers as signals for firm ESG (environmental, social, and governmental) efforts and high managerial ability (Welch and Yoon 2023); they find evidence that high-ability managers allocate resources to ESG in a way that enhances shareholder value. As possible extensions, a textual analysis on review content may be conducted in the context of ESG to extract and compare voluntary ESG disclosures provided by corporations with employee feedback on the companies' ESG practices.

Glassdoor may also be investigated in future studies for its effectiveness in reducing information asymmetry by examining how information from Glassdoor is utilized by different stakeholders. For example, auditors might use Glassdoor data to identify potential risk factors; investors and analysts can use this information to evaluate the information environment of a company and predict the performance of the company in the future. For companies, the dataset could also be used to examine intracompany information asymmetry and the cost associated with centralized decision making.

Finally, the sentiment (tone) and text metrics calculated based on employees' text review content can be used to conduct different accounting research. For example, existing research finds that management manipulates their disclosure sentiment to hide information or mislead investors according to firm performance (Huang et al. 2014); it is interesting to investigate whether employees also express different sentiments (positive, negative, etc.) in their reviews based on firm performance. Similarly, the uncertainty metrics capture the uncertainty and doubts in employee reviews. Researchers can use the uncertainty metric to investigate whether such uncertainty reflects the confusing messages employees received within companies or feelings employees have with the companies and whether such uncertainty can be related to the information environment of companies and have impact on analyst forecasting performance.

In summary, the R code and illustrative dataset presented in this paper demonstrate that Glassdoor can capture crucial information about employees' perception, company culture, and information environment. And the dataset provides a rich source of information for accounting researchers.

VI. CONCLUSIONS

The anonymous nature of Glassdoor allows employees to share their feelings openly and without repercussions; this is important, as the information provided in Glassdoor may be more accurate than other survey data. To better understand and use the employee review data provided in Glassdoor, this paper introduces different variables that can be extracted or calculated. This paper also provides the complete R codes for each step of data collection, cleaning, and variable construction processes.

We collected information from employees' ratings and reviews of S&P 1500 firms as of September 30, 2021. For each Glassdoor review, we collect three types of information: review metrics, review content, and reviewer information. Some popular textual metrics are also provided so researchers do not need to go through all the tedious and repetitive steps to calculate them. We believe our dataset can provide researchers with a quick start to evaluate their research ideas related to public employee reviews on social media platforms. If researchers need to refine the metrics or calculate those metrics based on new textual data, they can take advantage of our codes and stay focused on the researcher questions they are trying to study, not the laborious and time-consuming work on the commonly used metrics.

However, we admit that Glassdoor reviews might not represent the overall employee opinions and are subject to self-selection and manipulation. In other words, we can only observe the reviews by those employees who self-selected to post reviews/opinions on the Glassdoor platform. There could be many employees with different reviews/opinions but decide not to post them on the Glassdoor platform. Some reviews that contain useful information could be removed by Glassdoor, either because of policy violation (e.g., reviews with prohibited/restricted content) or other considerations. In addition, Glassdoor may foster a tendency toward negative bias, as reviews are submitted anonymously and there is little barrier for individuals to leave abusive or dishonest feedback. Unhappy current or former employees, who are often motivated by feelings of resentment, could post reviews with exaggerations and distortions. Even competitors could post negative reviews via fake accounts. Of course, there could also be some biased positive reviews intentionally posted by the company. Finally, the issue also arises as to whether Glassdoor opinion has a causal effect on economic outcomes if the two are correlated (Teoh 2018). For example, the firm may be experiencing some negative external shock, which caused employees' outlook to dim. In other words, it may not be employee outlook itself that affects firm performance but the external shock. In summary, our dataset has similar limitations of other review datasets from public websites, such as Facebook, Twitter (now X), Google Reviews, and Yelp. The dataset should be used with caution, and studies using our dataset should be conducted with appropriate research design and methods.

REFERENCES

Allee, K. D., and M. D. Deangelis. 2015. The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research* 53 (2): 241–274. https://doi.org/10.1111/1475-679X.12072

Babenko, I., and R. Sen. 2016. Do nonexecutive employees have valuable information? Evidence from employee stock purchase plans. *Management Science* 62 (7): 1878–1898. https://doi.org/10.1287/mnsc.2015.2226

Bright Data. 2024. How the world collects public web data. https://brightdata.com/

Brownlee, J. 2019. A gentle introduction to the bag-of-words model. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Chemmanur, T. J., H. Rajaiya, and J. Sheng. 2020. How does online employee ratings affect external firm financing? Evidence from Glassdoor. (Working paper). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3507695

Cooper, A. E., D. L. Diab, and K. M. Beeson. 2020. Why spelling errors matter: Online company reviews and organizational attraction. *Corporate Reputation Review* 23 (3): 160–169. https://doi.org/10.1057/s41299-019-00075-z

deHaan, E., N. Li, and F. S. Zhou. 2023. Financial reporting and employee job search. *Journal of Accounting Research* 61 (2): 571–617. https://doi.org/10.1111/1475-679X.12469

Dube, S., and C. Zhu. 2021. The disciplinary effect of social media: Evidence from firms' responses to Glassdoor reviews. *Journal of Accounting Research* 59 (5): 1783–1825. https://doi.org/10.1111/1475-679X.12393

Glassdoor. 2022. Making worklife better. https://www.glassdoor.com/about/

Glassdoor. 2024. How to Use Glassdoor to search for a job. https://www.glassdoor.com/blog/guide/how-to-use-glassdoor/

Green, T. C., R. Huang, Q. Wen, and D. Zhou. 2019. Crowdsourced employer reviews and stock returns. *Journal of Financial Economics* 134 (1): 236–251. https://doi.org/10.1016/j.jfineco.2019.03.012

Hales, J., J. R. Moon, Jr., and L. A. Swenson. 2018. A new era of voluntary disclosure? Empirical evidence on how employee postings on social media relate to future corporate disclosures. *Accounting, Organizations and Society* 68–69: 88–108. https://doi.org/10.1016/j.aos.2018.04.004

Hammami, A., R. Moldovan, and E. Peltier. 2020. Salary perception and career prospects in audit firms. *Managerial Auditing Journal* 35 (6): 759–793. https://doi.org/10.1108/MAJ-11-2019-2475

Hope, O. K., C. Li, A. P. Lin, and M. J. Rabier. 2021. Happy analysts. *Accounting, Organizations and Society* 90: 101199. https://doi.org/10.1016/j.aos.2020.101199

Huang, K., M. Li, and S. Markov. 2020. What do employees know? Evidence from a social media platform. *The Accounting Review* 95 (2): 199–226. https://doi.org/10.2308/accr-52519

Huang, M., P. Li, F. Meschke, and J. P. Guthrie. 2015. Family firms, employee satisfaction, and corporate performance. *Journal of Corporate Finance* 34: 108–127. https://doi.org/10.1016/j.jcorpfin.2015.08.002

Huang, M., A. Masli, F. Meschke, and J. P. Guthrie. 2017. Clients' workplace environment and corporate audits. *Auditing: A Journal of Practice & Theory* 36 (4): 89–113. https://doi.org/10.2308/ajpt-51691

Huang, X., S. H. Teoh, and Y. Zhang. 2014. Tone management. *The Accounting Review* 89 (3): 1083–1113. https://doi.org/10.2308/accr-50684

Huddart, S., and M. Lang. 2003. Information distribution within firms: Evidence from stock option exercises. *Journal of Accounting and Economics* 34 (1–3): 3–31. https://doi.org/10.1016/S0165-4101(02)00071-X

Jayan, J. 2023. Beyond basics: Advanced web scraping strategies for data professionals. https://www.promptcloud.com/blog/beyond-basics-advanced-web-scraping-strategies-for-data-professionals/

Karabarbounis, M., and S. Pinto. 2018. What can we learn from online wage postings? Evidence from Glassdoor. *Economic Quarterly* 104 (4): 173–189. https://doi.org/10.21144/eq1040402

Khavis, J. A., and J. Krishnan. 2021. Employee satisfaction and work-life balance in accounting firms and audit quality. *Auditing: A Journal of Practice & Theory* 40 (2): 161–192. https://doi.org/10.2308/AJPT-18-029

Khavis, J. A., J. Krishnan, and C. Tipton. 2022. Implications of employee satisfaction and work-life balance in accounting firms. *Current Issues in Auditing* 16 (1): P16–P26. https://doi.org/10.2308/CIIA-2021-006

Lee, Y., S. Ng, T. Shevlin, and A. Venkat. 2021. The effects of tax avoidance news on employee perceptions of managers and firms: Evidence from Glassdoor.com ratings. *The Accounting Review* 96 (3): 343–372. https://doi.org/10.2308/TAR-2019-0148

Lei, L., Y. Li, and Y. Luo. 2019. Production and dissemination of corporate information in social media: A review. *Journal of Accounting Literature* 42 (1): 29–43. https://doi.org/10.1016/j.acclit.2019.02.002

Li, F., M. Minnis, V. Nagar, and M. Rajan. 2014. Knowledge, compensation, and firm value: An empirical analysis of firm communication. *Journal of Accounting and Economics* 58 (1): 96–116. https://doi.org/10.1016/j.jacceco.2014.06.003

Li, M. 2019. Moral hazard and internal discipline: Theory and evidence. *The Accounting Review* 94 (4): 365–400. https://doi.org/10.2308/accr-52294

Loughran, T. I. M., and B. Mcdonald. 2014. Measuring readability in financial disclosures. *The Journal of Finance* 69 (4): 1643–1671. https://doi.org/10.1111/jofi.12162

Loughran, T. I. M., and B. Mcdonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54 (4): 1187–1230. https://doi.org/10.1111/1475-679X.12123

Saini, G. K., and I. M. Jawahar. 2019. The influence of employer rankings, employment experience, and employee characteristics on employer branding as an employer of choice. *Career Development International* 24 (7): 636–657. https://doi.org/10.1108/CDI-11-2018-0290

Scraping Robot. 2022. Scraping websites for academic research and school projects. https://scrapingrobot.com/blog/scraping-website-for-academic-research/

Shan, C., and D. Y. Tang. 2023. The value of employee satisfaction in disastrous times: Evidence from COVID-19. *Review of Finance* 27 (3): 1027–1076. https://doi.org/10.1093/rof/rfac055

Teoh, S. H. 2018. The promise and challenges of new datasets for accounting research. *Accounting, Organizations and Society* 68–69: 109–117. https://doi.org/10.1016/j.aos.2018.03.008

Welch, K., and A. Yoon. 2023. Do high-ability managers choose ESG projects that create shareholder value? Evidence from employee opinions. *Review of Accounting Studies* 28 (4): 2448–2475. https://doi.org/10.1007/s11142-022-09701-4

INDEX OF SUPPLEMENTAL MATERIALS

Please review the READ FIRST user agreement.
CodesAndDatasetInformation.docx
JIS_Glassdoor_review_data_sp1500_2021.rds
JIS_Glassdoor_review_data_sp1500_2021_metrics.rds
JIS_Glassdoor_review_data_sp1500_2021_sample.csv
JIS_Glassdoor_review_data_sp1500_2021_sample_metrics.csv
JIS_Glassdoor_Data.Rproj
JIS_Glassdoor_Data_R_Codes.Rmd
JIS_Glassdoor_Data_R_Codes_LocalDemo.Rmd
JIS_Glassdoor_Data_R_Codes_LocalDemo.html
JIS_Glassdoor_landingPages.csv
LM_Dic_1993_2021.rds
textualVars.R
reviews_14898_3D-Systems_1.html
reviews_14898_3D-Systems_2.html
reviews_14898_3D-Systems_3.html
reviews_65406_1800PetMeds_1.html
reviews_65406_1800PetMeds_2.html
JIS_Glassdoor_individualpage_urls.csv
JIS_Glassdoor_reviews_data_raw.csv
JIS_Glassdoor_review_data_clean.csv
JIS_Glassdoor_review_data_clean.rds
JIS_Glassdoor_review_data_text_metrics.csv
JIS_Glassdoor_review_data_text_metrics.rds
JIS_Glassdoor_review_data_covid.csv
JIS_Glassdoor_review_data_covid.rds