A Study on the Attributes of Companies, and their Effect on the Overall Rating on Glassdoor

Erin M. Swan-Siegel

Northwest Missouri State University, Maryville MO 64468, USA
S201660@nwmissouri.edu

Abstract. **Keywords:** business · employee satisfaction · company rating

1 Introduction

Every day, we are bombarded with products and businesses touting the product's utility or the benefits of their particular service through qualitative or quantitative summary statements; declarations such as "9 out of 10 doctors recommend", "4.5 stars with over twenty-thousand reviews!", and "Voted the best in the Midwest". We consumers rely on these ratings and reviews to help us decide on everything from where to eat, what product to buy, where to spend our vacation, and even where to work.

But what special mixture of characteristics contribute to a higher rating? Can we predict a company's rating based on key characteristics? Are there certain attributes that are closely associated with a 4-star rating versus a 1-star rating? Additionally, what business values are present in the company description when the Glassdoor rating is high?

1.1 Associated Literature

(Not completed) How is the Glassdoor rating determined? Is it past and present employees only or is it also those interviewed? Is there a bias there? Other similar studies?

1.2 Goals of this Study

The aim of this study was two-fold: Determine which attributes from available Glassdoor data are most closely related to the company's rating on the site in order to predict a company's rating when provided with five key characteristics; and what keywords are found in an employer's description more often in those with higher scores? To come to a conclusion on both matters, Machine Learning processes, relational models, and Natural Language Processing were employed, and the data was subsequently analyzed for meaning.

1.3 Methodology

Following the guidelines and advice of YouTuber Ken Jee[2], and heavily editing an existing Glassdoor web-scraper[3], a robust data-gathering script was written to create a data frame[4]. Substituting for the output of the script, due to scraping-prevention measures, a semi-structured dataset with eight attributes and nearly ten-thousand records was used.

In order to answer the questions posed by this study, a dataset containing five key metrics - Company Rating, Number of Company Reviews, Average Company Salary, Number of open jobs, and Number of Employees - was explored and cleaned. Records which did not contain all five attributes were removed from the machine learning training and test sets. Using SpaCy (NLP), keywords from the company description for the employers with a rating of 3.5 and above were determined.

Limitations to this study are that the Glassdoor rating found is not based on every past or present employee, just the subset who decided to participate. These reviews can also be affected by employees who are new to the company, or have recently been denied for a promotion. In summary, the context and attributes of those contributing to the rating cannot be accounted for or controlled, and must therefore be taken with the common-sense understanding that many consumers do for other location, service, or product page.

2 Data

Data for this study was intended to be gathered through a web-scraper built together with Software Engineer, Daniel Swan. The scraper extracts company data from an employer card found on a particular Glassdoor site and clicks into each company's Glassdoor employer page to extract additional data. Unfortunately for this study, companies are continuing to build better protections over their systems and their data, and the web scraper built for this study - while it does work as expected - encountered scraping prevention measures.

Web scraping prevention exists for many reasons, including blocking excess traffic that slows down the site for non-bot users and for protecting a company's intellectual property[1]. Options to bypass these security measures are accessible, but required significant re-work of the author's script. With a similar output resource available, data published to Kaggle in May 2023[3] was analyzed instead.

References

1. GPT, C.: Why websites prevent data scraping, https://chat.openai.com/
2. Jee, K.: Data science project from scratch, https://www.youtube.com/@KenJee_ds
3. Shil, J.: https://www.kaggle.com/datasets/joyshil0599/glassdoor-company-insightsscraped-data-collection/code
4. Swan-Siegel, E.: Web-scraper, https://github.com/progswan2022/scraping-glassdoor-selenium