Evaluating Stock Market Performance Using Aggregated Employee Reviews

Evaluating Stock Market Performance Using Aggregated Employee Reviews

an undergraduate thesis presented

by

Peter A. Ayala

to

The Department of Statistics

in partial fulfillment of the requirements

for the joint degree of

Bachelor of Arts

in the subjects of

Statistics and Computer Science

with Honors

Harvard University

Cambridge, Massachusetts

May 2019

Thesis advisor: Professor Michael Parzen Peter A. Ayala

Evaluating Stock Market Performance Using Aggregated Employee Reviews

Abstract

Investors are constantly driven by a desire to outperform the market. While investments may perform well in the short-term, there is great difficulty in achieving long-term success. We show that using aggregated employee reviews left on Glassdoor can result in significant performances spanning multiple years. A portfolio managed as the top quintile of firms by their Glassdoor rating has a significant Four-Factor alpha and excess returns above the S&P500. Strong positive changes in employee's perception on "Senior Leadership" and "Culture & Values" can also be used to create significant portfolio performance, with its significance additionally tested through a Fama-MacBeth regression. Lastly, using a text classification model to categorize employee reviews, we show that the dimensions "Agility," "Customer Orientation," and "Engagement" are indicative of significant positive performance.

Contents

1 Introduction 1

2 Background Information 5
2.1 Glassdoor Data . . . 6
2.2 Financial Performance . . . 10

3 Glassdoor Metrics 14
3.1 Portfolio Sorting Through Ratings . . . 15
3.2 Portfolio Sorting Through Recent Changes In Ratings . . . 18
3.3 Fama MacBeth Regression . . . 20

4 Analyzing Textual Responses 24
4.1 Textual Responses Summary . . . 25
4.2 Classification Methodology . . . 26
4.3 Determining Incidence and Sentiment of Values . . . 30
4.4 Analysis of Classification Output . . . 32
4.5 Panel Analysis Using Classification Output . . . 35
4.6 Portfolio Sorting Using Classification Output . . . 37
4.7 Clustering Issues in High Performing Firms . . . 38
4.8 Discussion . . . 39

5 Conclusion 40
5.1 Discussion . . . 40
5.2 Further Analysis . . . 42
5.3 Conclusion . . . 42

Appendix A Appendix 44

References 50

Listing of figures

2.1 Distribution of Firm Level Glassdoor Ratings 8
2.2 Glassdoor Individual Responses 9
2.3 Alpha Distribution 12

3.1 Glassdoor Metrics Correlations 15

4.1 Sentiment Correlations 34
4.2 Histograms of Sentiment Values 34
4.3 QQ plots of Sentiment Values 35

A.1 Incidence QQ Plots Pre Transformation 45
A.2 Incidence QQ Plots Post Transformation 45

List of Tables

2.1 Average Firm within Dataset by GICS Sector 7
2.2 Glassdoor Ratings Summary Statistics 8

3.1 Portfolio Returns Based on Ratings Over Time 16
3.2 Portfolio Returns Based on Short Term Changes in Ratings 18
3.3 Portfolio Returns Based on Short Term Changes in Subcategories 19
3.4 Fama-MacBeth Regression Results 22

4.1 Summary Statistics on Text Reviews 25
4.2 Description of Ten Value Classifications 26
4.3 Incidence Summary Statistics for Text Based Classification 31
4.4 Sentiment Summary Statistics for Text Based Classification 33
4.5 Fama MacBeth Regression Using All Classification Values 36
4.6 Individual Fama MacBeth Regressions Using Classification Values 37
4.7 Portfolio Sorting Based on Sentiments Results 38

Acknowledgments

There are numerous people whom I'd like to thank for making this thesis possible. First and foremost, I would like to thank the research team at MIT who made this research possible. My deepest gratitude goes to Donald Sull, who has acted as an advisor both within and outside this research since I began roughly a year and a half ago. To Andrei and Ryan, I thank for your continual and relentless work to refine the textual classification model. The results presented in this paper are in large part built on your shoulders. I also thank Glassdoor, whose data roots the results found in this thesis.

To all of my friends who observed me through this journey, I thank you for the constant words of encouragement and continual support offered. Opting to pursue a Joint and thesis so late into my college career was ambitious and could not have been done without the constant help reassurance received. Though there are too many to name, I'd especially like to thank Sherry Gao and Brian Marinelli for their ability to make everything seem right when at times everything felt wrong.

Last but not least, I would like to thank my parents and my brothers for their never-ending support and love both within and outside my time at Harvard.

1

Introduction

Public companies are consistently and rigorously analyzed by investors to determine signals that may indicate near term stock fluctuations. Large efforts have been spent analyzing the unique information certain employees may about about their companies. Notably, many papers find strong predictive power from top executive behavior, particularly at the corporate suite level[6,7,15,40,44]. These papers draw on significantly more publicly available in-

formation on executives than the common employee, and conclude actions by high ranking employees are indicative of future lawsuits[40]. This is further evidenced by papers relating negative press and media reports by executives to firm's near term financial performance[7]. However, the effect of knowledge lower ranking employees contain is not as clear as data is harder to gather. Though the research is less common, Guiso et al. find strong positive association relating an employee's perception of their firm's integrity to Tobin's Q, a popular measure of investment[28]. Edmands performed an aggregate portfolio analysis based on yearly "Great Places to Work" rankings and found significant portfolio returns above market rates, suggesting that the market doesn't fully capture more intangible information[19]. We conjecture that the aggregate view of employees left on Glassdoor reviews may be indicative of near and long term financial performance due to the unique information employees have and provide in their reviews.

This paper attempts to build on the success of previous papers by aggregating and analyzing information provided by the common employee. Specifically, we use the database of employee reviews available on Glassdoor. Glassdoor is a privately held company in which employees can leave anonymous and public reviews on their companies, and has garnered a total of over 5 million full-time employee reviews since its inception in June of 2007. The benefit of this database is its large size spanning multiple industries, geographic regions, and employee positions, which will help us generalize conclusions across a larger area of interest. The granularity of having individual employee reviews for each firm allows us to greater differentiate specific culture values between the firms though a textual classification model.

The work presented in this thesis augments results found in other research streams. Under the efficient market hypothesis, we'd expect there no abnormal relationship to exist

between the publicly available information on Glassdoor and market returns. If a relationship were to exist, then this would suggest a market inefficiency relating corporate ratings to abnormal returns. Using a portfolio sorting strategy to reflect changes in ratings at various times, we find significant alpha returns for high-rating portfolios, consistent with Edman's major findings when performing a similar study using the "100 Best Places to Work Dataset."[19]

We also conjecture that recent changes in employee perceptions about their company can signal near term predictions about their returns. This follows from the theory that employees at a firm would have a general sense of changes within the company before market does. We regress quarterly returns against a lagged shift in changes in ratings, but find generally weaker results than using the firm's rating at a given time. However, we do find significant alpha deviations when considering strong positive changes in the perception of "Career Opportunities," "Senior Leadership," and "Culture & Values," suggesting similar findings to those that claim senior management behavior is indicative of future firm behavior or the strong impact of having a positive and well-loved culture[15,36,37,45].

Lastly, we augment the above by analyzing the textual responses left by employees. By classifying reviews both by their sentiment and incidence across various topics that might be written about in a review, we find significant predictors relating culture values and returns. Specifically, we find that the values "Agility", "Engagement", and "Customer Orientation" to be significant in predicting alpha values when used to select portfolios, supporting research relating these key values to long term performance[28,36].

We structure this thesis into three main chapters. Chapter 2 discusses additional background information about the various data sources, rationale for performance metrics, gen-

eral summary statistics, and definitions for relevant terms. Chapter 3 explores the relationship between Glassdoor firm ratings and stock performance as measured through returns and alpha values. Chapter 4 explores the textual responses users leave through a classification model and the resulting predictive power for firm performance. The conclusion offers final remarks about the results presented in this thesis.

We live in an age awash with information. Readers don't just want random snatches of information flying at them from out of the ether.

Jack Fuller

2

Background Information

For the purposes of have a greater understanding for the rest of the paper, we present a more rigorous introduction on the background of this project.

2.1 Glassdoor Data

The data is provided by Glassdoor, a privately held employer review and recruiting website created in 2008. Employees, both former and current, are able to leave anonymous reviews on their companies about their work-life balance, compensation, career opportunities, and the likes. In order to leave a review, the employee must confirm their employment through a verification email sent to their work email. Users are encouraged to leave reviews on their employers in exchange for increased access to different parts of the website, as well as contributing towards a common good of complete information. Glassdoor actively monitors reviews written to remove obvious outliers in both the positive and negative directions, thereby making the resulting dataset more indicative of a firm's true rating.

A composite Glassdoor review is composed of both mandatory and optional responses. Each employee review must have a one to five rating on the *Overall Rating* of the company, and an optional one to five rating for *Career Opportunities, Compensation & Benefits, Senior Leadership, Work/Life Balance*, and *Culture & Values*. Though the latter five are optional, each has roughly an 87% response rate as shown in Table 2.2. Additionally, users are required to submit textual responses to both *Pros* and *Cons*, though there is no character or word minimum. Lastly, users can leave an optional third textual response for *Feedback*, which has a response rate of 59.4%. There are a few other variables, some of which are imputed by Glassdoor like *Review ID* or *DateTime*, and others which are optionally submitted by the user, including *birthYear* and *Education*.

Control variables are also added to the dataset by querying information from both Compustat and CRSP. Each company has a corresponding sector assigned matching the Global

GICS Sector	# Firms	Log Employees (Thousands) (Mean)	Market Value (M) ($)(Mean)	# Reviews (Mean)
Consumer Discretionary	67	3.90	37,207	3,249
Consumer Staples	24	4.16	65,126	2,962
Energy	10	3.62	97,644	1,035
Financials	49	3.49	63,411	2,188
Health Care	54	3.66	56,057	1,381
Industrial	55	3.80	35,272	1,428
Information Technology	84	3.07	60,948	1,691
Materials	5	3.63	36,649	1,080
Telecom. Services	9	3.54	63,438	3,902

Table 2.1: In this Table we present summary statistics on the firms within the dataset of interest.

Industry Classification Standard (GICS). This standard is built using 11 sectors across 69 industries, though only 9 sectors are represented in the subsetted data. We also includes controls for industry size, measured as the log value of employees, and market value of each company, with averages shown in Table 2.1.

Though the company has acquired over five million reviews since its inception, we subset the dataset to one of greater interest. In particular, only full-time employee reviews are considered as they will be able to speak more accurately about the firm than part-time or intern employees[14]. We further subset companies to those that are publicly traded as their financial information is more readily available. Lastly, we only consider companies that have more than 200 reviews listed so that firms consistently have a large, accurate sample across periods of time. The resulting dataset consists of 361 companies containing a total of 774,738 reviews. We show the overarching summary statistics for the most pertinent variables in Table 2.2.

A concern when dealing with data submitted optionally by users is a bias towards polar ends of the spectrum. Initially, it may seem that users opting to write a review would

Variable	Percent Response (%)	Mean	Std. Dev.	Min	Max
Overall Rating	100	3.37	0.37	2.37	4.54
Career Opportunities	87.7	3.16	0.35	1.99	4.39
Compensation & Benefits	87.7	3.40	0.44	2.18	4.67
Senior Leadership	86.4	2.91	0.38	2.02	4.40
Work/Life Balance	87.8	3.26	0.39	2.19	4.49
Culture & Values	87.0	3.31	0.45	2.12	4.68

Table 2.2: This table reports the summary statistics for the Glassdoor defined variables, aggregated at the company level for firms within our subset dataset. The exception is "Percent Response," which reports the percentage of individual users that opted to fill in that particular field.

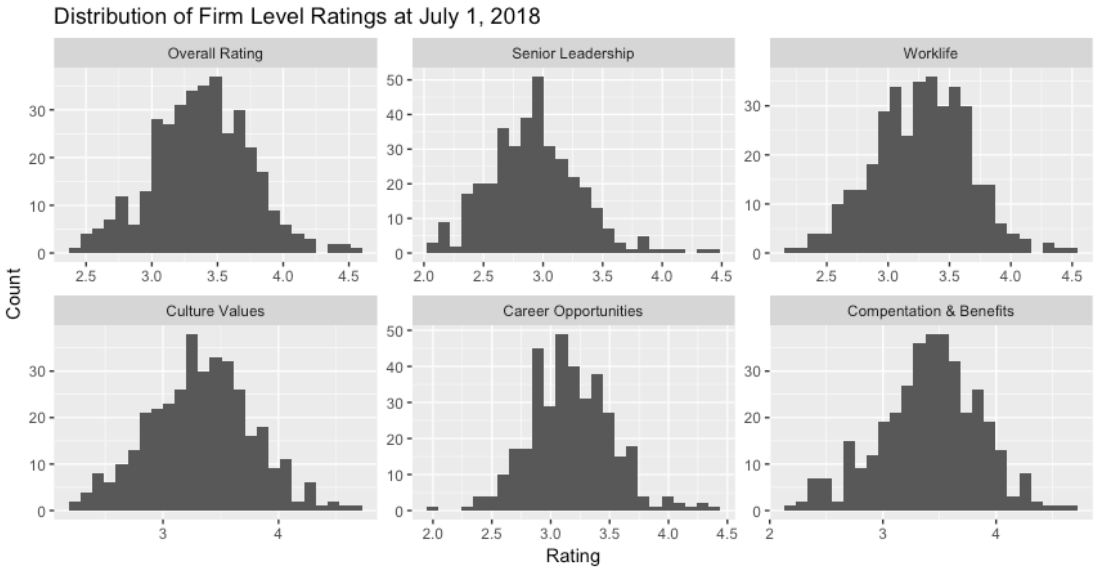

Figure 2.1: This figure shows the distribution of the ratings for the six metrics Glassdoor offers users, aggregated at the firm level on July 1, 2019. We see that all ratings are approximately normal.

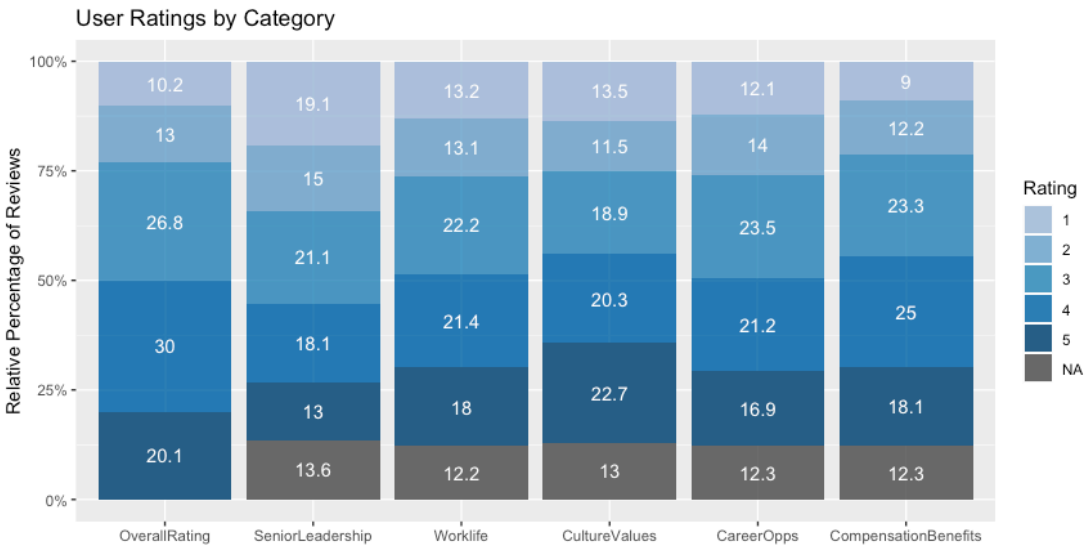

Figure 2.2: The figure shows the distribution of ratings by individuals. Particularly, there doesn't appear to be extreme amount of polar opinions by user

do so because they exist on some extreme in their thoughts about the company. However, plotting and reviewing the spread of responses by subcategory does not reveal an extreme amount of skew in either direction, as seen in Figure 2.2. Employees generally have an incentive to provide an honest evaluation about their company for the common good[39].

Because we opt to use only publicly traded companies, financial information is readily available through Compustat and The Center for Research in Security Prices (CRSP). The Beta Suite by Wharton Research Data Services was also queried, which aggregates predefined financial models using data from CRSP and Compustat as well. This data can be queried on a daily, weekly, or monthly, though for the purposes of this paper we look to monthly reports.

2.2 Financial Performance

Though there are many natural measures to describe financial performance, we start by looking at Jensen's Alpha, described as the rate of return above or below what is predicted through investors in the Capital Asset Pricing Model (CAPM)[8]. Specifically, the CAPM model describes

$$\mathbb{E}[R_i - R_f] = \alpha + \beta_i(\mathbb{E}[R_m] - R_f) + \varepsilon_i$$

where R_i represents the return of investment, R_f represents the risk-free rate of high yield savings or treasury bonds, and R_m describes the return of the market. β is formally defined as

$$\beta_i = \frac{\text{Cov}(R_i, R_m)}{\text{Var}(R_m)}$$

and generally represents the volatility of the stock relative to the market. Thus, any positive α value indicates a stock or portfolio return above what is already expected from its increased risk, and similarly any negative α value indicates an investment return below what should be expected given the risk. Under the efficient market hypothesis, we expect α to be zero to reflect the market's full understanding of the stock given all available information[24].

While the CAPM model does offer a good starting point on a firm or portfolio's performance above or below the norm, more sophisticated models have emerged. Specifically, the Fama-French three-factor model incorporates the knowledge that both small firms and firms with a high book-to-market ratio tended to perform better than market benchmarks[23]. The Fama-French model is described as

$$\mathbb{E}[R_i - R_f] = \alpha + \beta_i(\mathbb{E}[R_m] - R_f) + \beta_s SMB + \beta_v HML + \varepsilon_i$$

where SMB stands for "Small market valuations Minus Big" and HML stands for "High Minus Low." SMB captures the historic excess returns of small stocks over big stocks, and HML captures value stocks over growth stocks. Lastly, the Cahart Four Factor Model expands the Fama-French model by incorporating a "momemtum" term, which flows from the idea that firms that have historically done well will have momenta to continue upwards, and likewise firms that experience long periods of decline will continue their momentum downwards[10]. The Cahart Model is described by

$$\mathbb{E}[R_i - R_f] = \alpha + \beta_i(\mathbb{E}[R_m] - R_f) + \beta_s SMB + \beta_v HML + \beta_m WML + \varepsilon_i$$

where WML describes "Winners Minus Losers." That is, portfolio managers will tend to long stocks with consistent high momentum and short stocks that have low momentum. We use α described in the Cahart Four Factor Model, as the overall model evaluation captures about 90% of market variance and is shown to be most accurate[10,22]. Thus, any significant α findings will have more merit as there is less chance of unknown confounders. We estimate α by regressing against the returns for a given time period, and determine the significance level of the resulting intercept term. We query an individual's betas $\beta_i, \beta_s, \beta_v$ from Wharton Research Data Service's "Beta Suite" database on a monthly basis. We query values for WML, SMB, HML, R_f, and R_m from the "Fama-French Portofilio and Factors" database on a monthly basis as well, where the risk free return is the one-month treasury bond yield for a given month and the excess market return is the value-weighted portfolio

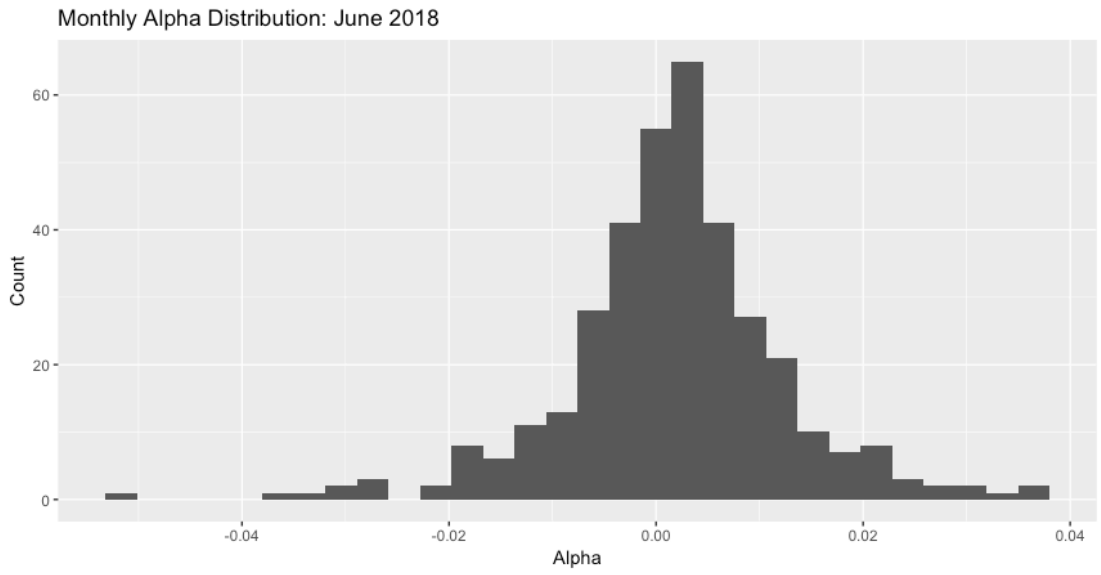

Figure 2.3: This histogram shows the distribution of monthly alphas for the companies in our dataset. The values are represented as the decimal versions of percents

of all stocks in the NYSE, AMEX, or NASDAQ minus the risk free return. The resulting distribution of an average monthly alpha at the firm level is shown to be approximately normal with mean 0.00176 and standard deviation 0.0107, as seen in Figure 2.3.

Lastly, we also formally define terms that may be used throughout the rest of the paper.

- Corpus: A corpus C represents the entire collection of written texts. In this paper, the corpus represents the entire Glassdoor database of employee's written reviews.
- Document: A document d represents a single text instance in the corpus such that $d \in C$. In this paper, a document refers to a single instance of a response in either the *Pros*, *Cons*, or *Feedback* fields.
- Review: A review by a user includes the collection of Glassdoor one-to-five metrics, and their responses to all of the fields in *Pros*, *Cons*, and *Feedback*.
- Pros: A textual input users are required to submit, with a guiding question of "Share

some of the best reasons to work at ..."

- Cons: A textual input users are required to submit, with a guiding question of "Share some of the downsides of working at ..."
- Feedback: An optional response with no guiding question, but recently changed from "Feedback" to "Advice to Management."
- Risk Free Return R_f: The return rate for a "zero"-risk investment. For the purposes of this paper, the one-month U.S. treasury interest rate for a given month t is the risk-free return.
- Return: The return R_i represents the return rate for a given investment.
- Excess Return: The return of an investment above or below the risk-free rate, calculated as $R_i - R_f$.
- Idosyncratic Volatility: The factors that affect an asset at the individual microeconomic level. We include this as a control when performing the Fama MacBeth Regression, queried for each firm on a monthly period through the Wharton Beta Suite.

Being a data scientist is not only about data crunching. It's about understanding the business challenge, creating some valuable actionable insights to the data, and communicating their findings to the business.

Jean-Paul Isson

3

Glassdoor Metrics

We begin our analysis by first looking at the one-to-five metrics employees leave on their companies within their reviews. Specifically, the variables of interest are *Overall Rating, Career Opportunities, Compensation & Benefits, Senior Leadership, Work/Life Balance*, and *Culture & Values*, where *Overall Rating* is the only variable that is mandatory.

The values are highly correlated with each other, showcasing that users leave similar ratings across all the subcategories. When looking at the aggregate firm level, the problem becomes more enunciated, as seen in figure 3.1. This causes multiple issues with multicollinearity when attempting to build traditional linear models.

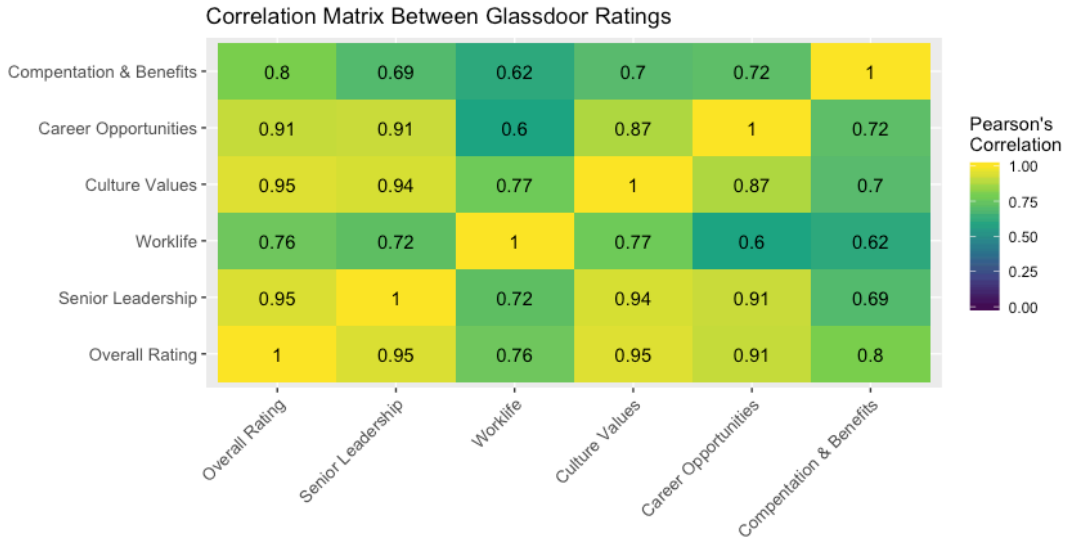

Figure 3.1: This heatmap shows the correlation between Glassdoor's one-to-five ratings. All metrics are strongly and positively correlated with each other, with the lowest pairwise correlation belonging to "Worklife" and "Compensation Benefits" at 0.6

3.1 Portfolio Sorting Through Ratings

As an alternative, we can take a portfolio sorting approach to determine if there might exist some relationship between ratings and returns. To assess whether on the whole a portfolio of a certain type may outperform the market generally, we categorize firms into sorted portfolios based on their rankings. First, we consider the hypothesis that high ranking companies perform better than the norm. To test this theory, we sort the reviews by the time

of their submission and compute the *Overall Rating* score at each calendar quarter for each company in the dataset. Starting from January 1, 2014, we construct three separate portfolios. A "High Rating" portfolio represents the top 20% of companies by *Overall Rating* at a given time, a "Normal Rating" portfolio represents the inner 60% of companies by *Overall Rating*, and lastly a "Low Rating" portfolio representing the lowest 20% of companies by *OverallRating*. This portfolio sorting division is consistent with those found in various other papers[20,19]. At each quarter Q_t, we calculate monthly α and return values for the portfolios created at Q_{t-1}, and re balance each portfolio according to an updated sorted list of firms by *Overall Rating*. Under our null hypothesis, we expect each of the three portfolios to have the same α values of zero, as readily available information should be accounted for in evaluations of a stock's price. Results are shown in Table 3.1

Portfolio	E. Return (%)	α (%)	t-stat	p-value
Low Rating	0.55	-0.20	-1.87	0.062
Medium Rating	0.88	0.11	1.22	0.221
High Rating	1.12	0.42	3.38	0.000
Market	0.93			

Table 3.1: The results table presents the findings from the quarterly managed portfolios between January 1, 2014 and July 1, 2018. Excess Returns and Alpha coefficients are given in monthly terms.

One of the major assumptions about the usual *t*-statistic is independence across measures. For time series data, and especially data about firm prices, this is usually broken as high performing stocks will likely be high performing for long periods of time, while low performing stocks will remain low performing, consistent with "momentum" of the stock typically observed. Thus, we use the Newey-West *t*-statistic to report significance, which uses a heteroskedasticity and autocorrelation consistent covariance matrix[48,42].

We see from Table 3.1 some pretty striking findings. Both the alpha and excess return

values for the portfolios monotonically increase across the portfolio ratings. We also see significant returns from the "High Rating" portfolio and significant α, indicating that the excess return above the market was abnormal given the risk taken on. Though the findings may seem egregious, these findings mimic findings published by Edmans, who used the "100 Best Companies to Work For in America" dataset to find that the market undervalues employee ratings[19]. Similar to Glassdoor, "100 Best Companies to Work For in America" releases yearly aggregate rankings on firms, which Edmans performed his analysis on. Specifically, his portfolio had a robust monthly Four-Factor alpha of 0.29% between 1984 and 2009 when considering all companies appearing on the "100 Best Companies" list, close to the monthly alpha found here.

It should be noted the above portfolios do not necessarily take into account large shifts that may occur within a company in a short period of time. For example, a company may be rooted in as having a high ranking overall from a long history of reviews, but are experiencing a recent downward trend due to recent changes within the company. Thus, we follow a similar procedure as above but now consider the change (Δ) of the *Overall Rating*, Δ*OverallRating*. Specifically, we compute Δ*OverallRating* for some time period Q_t as the average *OverallRating* at quarter Q_t minus the average *OverallRating* at quarter Q_{t-1}. To reduce the amount of variance caused by few number of responses, we only include companies that have more than 45 reviews in both the previous quarter and the current quarter. Similarly, we create three equally-weighted portfolios and compute alpha values for some time Q_t based on the Δ*OverallRating* at time Q_{t-1}. As Results for these portfolios are given in Table 3.2

Similar to the previous model, we report Newey-West *t* statistics to determine signifi-

Portfolio	Return (%)	α (%)	t-stat	p-value	Δ Mean
Low Δ Return	0.51	-0.13	-0.79	0.43	-0.29
Medium Δ Return	0.74	0.08	0.69	0.49	0.08
High Δ Return	0.97	0.19	1.69	0.092	0.31

Table 3.2: This table shows the monthly returns and alphas for the different portfolios, with both returns and alphas given in a monthly percentage format. None are significant at the $p = 0.05$ level, but the High Δ portfolio does just become significant at the $p = 0.10$ level

cance. We see that the only moderate significant result is the portfolio of High Δ Return at a p cutoff of 0.1, with neither the medium or low Δ resulting in significant t-statistics. This result may suggest firms experiencing very large increases in employee satisfaction in the short term signal positive alpha values in the near future. Similar to our findings from the first model, both the returns and associated alpha values are monotonically increasing through the portfolios.

3.2 Portfolio Sorting Through Recent Changes In Ratings

The above models are based on a portfolio selection driven by *OverallRating* with the thinking that it would be most indicative of an employee's overarching thoughts about their company. We now look to the other five metrics employees can opt to report. While *OverallRating* may act as the "umbrella" metric to the other variables, it is possible some shifts in subcategories may serve as better indicators of near term performance. For example, significant changes in *Senior Management* may serve as a stronger indicator of an employee's willingness to be productive for the company. We follow a similar model procedure as above, regressing quarterly returns against the sorted Δ of a particular value from the previous quarter. Results are shown in Table 3.3.

It's important to discuss the interpretability of each portfolio type. Low Δ firms char-

	Portfolio	Return (%)	α (%)	t-stat	p-value	Δ Mean
Career Opportunities	Low Δ	0.56	-0.09	-0.48	0.630	-0.32
	Med. Δ	0.92	0.17	1.80	0.073	0.00
	High Δ	1.02	0.30	2.41	0.016	0.34
Compensation & Benefits	Low Δ	0.83	0.08	0.56	0.577	-0.28
	Med. Δ	0.84	0.17	1.51	0.132	0.00
	High Δ	0.92	0.13	0.99	0.321	0.28
Senior Leadership	Low Δ	0.52	-0.16	-0.88	0.378	-0.35
	Med. Δ	0.89	0.19	2.03	0.042	0.01
	High Δ	1.11	0.35	2.85	0.004	0.36
Worklife	Low Δ	0.60	-0.14	-0.95	0.341	-0.33
	Med. Δ	0.95	0.23	2.06	0.040	0.00
	High Δ	0.83	0.20	1.95	0.052	0.34
Culture & Values	Low Δ	0.55	-0.17	-1.03	0.303	-0.34
	Med. Δ	0.92	0.18	2.00	0.046	0.01
	High Δ	1.01	0.32	3.02	0.003	0.36
Overall Market		0.93				

Table 3.3: We show the results of portfolio sorting by the various metrics Glassdoor defines. These metrics are optional for users to populate, though the majority do opt to respond. Alpha and excess returns for the portfolio are given in monthly percentage rates. See Table 2.2 for response rates by category.

acterize firms that experience sizable decreases in a particular subcategory, and similarly High Δ firms characterize firms that experience a sizable increase in a particular subcategory. Following from the statistical idea of "reversion to the mean," we'd generally expect High Δ firms to come from previously low ratings, and Low Δ firms to generally come from previously high ratings[29]. As a result, we may expect chosen firms to then undergo significant α changes, as the market may not be fully responsive to new changes within the ratings. However, the interpretability becomes more difficult when considering the Medium Δ portfolios. As we see from Table 3.3, these portfolios are centered at zero and thereby cause greater difficulty in determining what type of firm is represented. For example, firms that receive consistent rating scores, either positive or negative, will experience

limited Δ change and thus fall into the Medium Δ portfolio. As we saw in Table 3.1, significance can be achieved by choosing high performing firms consistently, and thus may result in significant alphas for Medium Δ portfolios as well. We potentially see this effect occur with *Worklife*, whose Medium Δ portfolio has a significant alpha while neither its High or Low portfolios are significnat.

With the exception of *WorkLife*, we see monotonically increasing portfolio returns for each subcategory. We also witness significant alpha values for High Δ portfolios in the subcategories of *Career Opportunities*, *Senior Leadership*, and *Culture & Values*. High perceived Δ shifts for *Career Opportunities* may reflect legitimate increases in a firm's near term growth prospects, potentially before the market fully reacts[45]. Senior Leadership is shown to be a strong determinant in firm success, and an employee's unique inside knowledge working directly under new management may be indicative of near term growth[12]. Lastly, large increases in *Culture & Values* may generally improve productivity and morale within the firm and therefore result in abnormal increases[31].

3.3 Fama MacBeth Regression

The portfolio method is valuable as it can offer guidance in determining significance based on a single metric, but can be open to unknown confounding variables. Because we're unable to control for other variables during portfolio selection, it might be the case that portfolios are being selected through some underlying unknown factor that influence ratings. We now look to see if we can replicate similar results while accounting for controls. To do so, we perform a cross-sectional regression to determine significance of our variable Δ*OverallRating*. In particular, we perform a Fama-MacBeth two-step Panel regression as a

way to explain returns while including controls for other factors.

The goal of the Fama-MacBeth is to find the return premium from the included factors. Each firm i in the time period experiences in own time series of the form

$$R_{i,t} = \alpha_i + \beta_{i,F_1} F_{1,t} + \beta_{i,F_2} F_{2,t} + \cdots + \beta_{i,F_m} F_{m,t} + \varepsilon_{i,t}$$

where $F_{j,t}$ is some included factor j at time t, and $R_{i,t}$ represents the excess return of firm i at time t, and B_{i,F_m} is the factor exposure. We then compute the T cross-sectional regressions of the returns on our estimates, $\hat{\beta}$. That is

$$R_{i,t} = \gamma_{t,0} + \gamma_{t,1} \hat{\beta}_{i,F_1} + \gamma_{t,2} \hat{\beta}_{i,F_2} + \cdots + \gamma_{t,m} \hat{\beta}_{i,F_m} + \varepsilon_{i,t}$$

The γ are then considered the risk premium for each factor. In the case of our work, we describe the cross-sectional regressions of monthly returns against the lagged quarterly changes in ratings and other firm constants. We have

$$R_{i,t+1} = \gamma_{0,t} + \gamma_{1,t} \Delta\mathrm{Metric}_{i,t} + \gamma_{i,t} \mathrm{X}_{i,t} + \varepsilon_{i,t}$$

where $R_{i,t+1}$ represents the excess return for firm i and time $t+1$, $\Delta\mathrm{Metric}_{i,t}$ represents the most recent change in quarter-by-quarter metric at time t, and $\mathrm{X}_{i,t}$ represents the vector of controls for firm i at time t. We use as controls the idiosyncratic volatility of each stock at month t, total volatility of each stock at month t, the excess return at time R_{t-1}, the averaged monthly excess returns between $R_{t-12:t-2}$, and the number of employees within the company .

Performing the Fama-MacBeth regression using $\Delta Metric = \Delta OverallRating$ results in similar findings as the general portfolio division. The resulting γ estimate has a Newey-West adjusted t-statistic of 1.49, resulting in a p-value of 0.146, as shown in 3.4. Similarly, we perform individual Fama-MacBeth regressions using $\Delta Metric = \Delta SeniorLeadership$, $\Delta Metric = \Delta Culture\&Values$, and $\Delta Metric = \Delta CareerOpportunities$. As we see in the resulting tables, both *Senior Leadership* and *Culture & Values* remain significant in the regression when accounting for controls, suggesting that there exists positive relationships between these deltas and near term excess returns.

	$\Delta O.R.$	$\Delta S.L.$	$\Delta C.V.$	$\Delta C.O.$
Δ Metric	1.45	2.36	4.02	1.64
Constant	1.23	0.94	0.94	0.93
Idy. vol	2.01	1.35	1.34	1.35
Tot. Vol.	-2.15	-1.55	-1.54	-1.54
R_{t-1}	-3.71	-3.15	-3.14	-3.14
$R_{t-12:t-2}$	0.88	1.03	1.00	1.04

Table 3.4: We present the t-statistics for the four Fama MacBeth regressions performed on the quarterly deltas of Overall Rating, Senior Leadership, Culture & Values, and Career Opportunities. Senior Leadership and Culture Values are both significant in the model accounting for controls.

The results presented in this chapter are consistent with similar research, yet still add nuanced conclusions to support. Performing portfolio division based solely on the *Overall Rating* of the company at a given time presents slightly inflated alpha values relative to those presented by Edmans using the "Great Places to Work" dataset, but may also be a result of differences in time periods the analysis was performed on. When considering short term changes in a firm's categories, we show that increases in the deltas for *Senior Leadership* and *Culture & Values* can be predictive of near-term excess returns and alpha performance. This augments claims by researchers relating the importance of strong Senior

Leadership and financial performance[37]. *Culture & Values* is particularly interesting as it encapsultes various measures, but its deltas in the positive may signal future growth within the company due to happier employees[36].

He writes the worst English that I have ever encountered. It reminds me of a string of wet sponges; it reminds me of tattered washing on the line; it reminds me of stale bean soup, of college yells, of dogs barking idiotically through endless nights. It is so bad that a sort of grandeur creeps into it. (Writing about US President Warren G. Harding)

Henry Louis Mencken

4

Analyzing Textual Responses

We now look to a new model that seeks to provide further granularity to the results presented earlier. The previous chapter focused in great detail on the one-to-five ratings left by reviewers in Glassdoor's pre-defined metrics, but now we look to the textual responses left in the *Pros, Cons*, and *Feedback* section of the reviews.

4.1 Textual Responses Summary

To provide further granularity on the the positive and negative ratings given in the Glassdoor metrics, we look to build a classification model that captures the various culture traits reference in these reviews. Factors including the variability in response length, writing style, and topic focus areas makes building a classification model difficult. Some employees will opt to write in prose format, while others opt to provide a bulleted list of incomplete sentences. Most reviews sit at only a few sentences long, with the longest average review going to *Cons* at only 29 words. These short responses, in combination with no clean supervised classification dataset to train on, make traditional classification algorithms difficult to successfully use. Brief summary statistics for the length of reviews is given by Table 4.1.

Variable	Response Rate (%)	Mean	Std. Dev.	Q1	Q3	Max
Pros	100	18.1	21.2	7	22	1190
Cons	100	29.2	49.3	8	31	2637
Feedback	59.4	13.6	25.3	1	18	5112

Table 4.1: This table shows the summary statistics on individual review lengths as denoted by the number of words. Q1 and Q3 describe Quartile 1 and Quartile 3, respectively.

Specifically, we hope to classify the reviews based on values both employees and firms deem important to a company's longevity. These values are defined as *Agility, Collaboration, Customer Orientation, Engagement, Execution, Inclusivity, Innovation, Integrity, Performance*, and *Respect*, and capture various nuances in company culture. Detailed descriptions of each of these values are given in Table 4.2, and is the result of managerial economic research and mimic closely what employer's themselves advertise as important traits about their company[28].

Value	Description
Agility	Speed to which an organization enacts new policies or reacts to shifts
Collaboration	How well different teams and departments cooperate with one another
Customer Orientation	Extent to which an organization focuses on customers and meeting their needs
Engagement	How involved employees are, both in their work and feedback
Execution	Extent to which projects are carried out successfully
Inclusivity	Extent to which an organization offers a welcoming environment
Innovation	Extent to which the organization improves itself and takes risks
Integrity	Employee's views on unethical, dishonest, or unfair policies within the company
Performance	Extent to which employees feel satisfied that performance is accurately promoted and evaluated
Respect	Extent to which employees and managers treat each other with respect

Table 4.2: We present the 10 large classification values of interest and their definitions.

4.2 Classification Methodology

The simplest model of feature generation is to use Bag-of-Words, where a document is represented as a vector of word counts mapped against the entire corpus dictionary. At its most basic, Bag-Of-Words uses single words, also known as one-grams, as its keys. The corpus dictionary can be improved by including higher order grams, though the process becomes more computationally expensive and the resulting vector space becomes more sparse. Using Bag-of-Words alone is often insufficient and rudimentary, as word counts do not take into account sentence structure or sentence lengths, but does lay the foundation for various other methods widely used[34].

Latent Dirichlet Allocation (LDA), for example, is a popular unsupervised topic modeling algorithm that clusters relevant documents together based on individual word groupings within documents[9]. While this helps determine certain clusters among your docu-

ments, it may not naturally converge towards topics one may specifically be interested in. This is further problematic in the scenario in which the documents or topics are not completely unique, which is the case of these reviews. GuidedLDA is a recent modification to the original algorithm that attempts to alleviate these problems by allowing you to specify both the topics of interest and seed words that belong to that cluster[47]. While this is a step in the right direction, it still remains difficult to sort a textual instance into multiple classifications. In both LDA and GuidedLDA, the output is a vector of probabilities assigning a word to a specific classification, so it becomes difficult to convert this to a multi-classification output.

Thus, we look to build a varied classification model that supports the ability to have multiple classifications for a single review. Specifically, we take a keyword and phrase approach to this problem similar to GuidedLDA, where each classification grouping of interest has an associated dictionary of words and phrases that would be indicative of the presence of this particular grouping. For example, we might take the word "fast" to be indicative that the employee is discussing the speed at which the firm operates. However, looking at the root of the word is insufficient. Take, for example, the following four sentences.

- The fast paced environment means you're constantly learning.
- On the whole I wish the company moved faster
- With the amount of meals you miss due to meetings you may as well be fasting
- It's fast food - what do you expect

The sentences highlight some key difficulties with simply using words as a way to classify sentences. In the first, the word fast is used in a way that is truly indicative of a company's general speed of operation. The second sentence showcases how we must also consider

the different tenses of a particular key word to capture all instances. However, the third sentence emphasizes that considering all potential tenses may result in differences in definitions. The last sentence emphasizes that the presence of a key word without any tense modifications still may not be indicative of a classification, as the phrase it is in can mean something different.

To account for these situations, the majority of indicative components in the dictionary are phrases, and are also paired with exclusion terms. For example, the word ”food” would be an exclusion term for the term ”fast” when trying to map some review to the classification Agility, so that the phrase ”fast food” does not result in a false positive. Furthermore, key phrases added into the dictionary take into account differences in where pairs might be located in a sentence. For example, the pair ”work∼life” is able to capture the phrases ”work life,” ”work my life away,” ”work my whole life away,” ”work my whole entire life away,” or any length of n-grams between ”work” and ”life.” Specifically, the output classification of an entire review is the concatenation of three separate vectors describing the output for *Pros*, *Cons*, and *Feedback*. This can be described as

$$\vec{R_i} = [\vec{P_i}, \vec{C_i}, \vec{F_i}]$$
$$\vec{P_i} = [s_0, \ldots, s_j] \text{ for } j \in \{\text{Values}\}$$
$$\vec{C_i} = [s_0, \ldots, s_j] \text{ for } j \in \{\text{Values}\}$$
$$\vec{F_i} = [s_0, \ldots, s_j] \text{ for } j \in \{\text{Values}\}$$

where R_i represents the overall review for an individual containing elements Pros classification P_i, Cons classification C_i, and Feedback classification F_i, with s_j denoting the sum

number of matches between words or phrases existing in a review and its existence in the key dictionary.

A general difficulty with building a classification model this way is that the classification success is based on the key-term dictionary you provide it. This requires a significant amount of manual effort to create, improve, and regulate new keys added and their resulting classifications. While it may be easy to create key words to define a new topic in the beginning, it becomes difficult to account for all nuances that may exist across all reviews. Therefore, manual inspection on classification outputs for new terms added is required to reduce the number of false positives created. However, the benefit in building a classification model of this sort is complete control in what phrases classify into what topics, as well as a consistent, fully deterministic output. Once a particular topic has a sizable number of key phrases as well, it becomes easier to determine potential unknown words by running a GuidedLDA on that particular topic. In it's current form, the dictionary has a total of 5,910 key terms with associated exlcusion terms as well.

Textual reviews are given as raw strings by Glassdoor, so general pre-processing is first applied to clean and standardize all the reviews. This includes the removal of line breaks denoted by "\n" or "\r", parentheses, forward or backslashes, and a few other edge cases. Reviews are then scanned by comparing words and phrases against the dictionary of key phrases, and classifications are outputted and separated by a user's *Pros*, *Cons*, and *Feedback* fields.

There are two different metrics that we want to capture when dealing with textual classifications, which are incidence and sentiment. Incidence needs to capture the rate at which employees talk about a particular topic within a company, while sentiment needs to capture

whether employees within a firm are describing this topic in a positive or negative sense.

4.3 Determining Incidence and Sentiment of Values

Incidence is the easier of the two metrics to capture. We define the incidence of a value for a particular firm as the proportion of employees that opted to write about that value in either the *Pros*, *Cons*, or *Feedback* field in their review. The classification output for a particular review is the sum counts of matches between reviews phrases and key phrases in the dictionary. Any positive number for a particular topic for a given review is converted to a one, indicating that the employee opted to talk about that particular value in at least one of the three textual responses. This transformation allows us to treat each individual reviewer equally so that lengthy or extremely passionate reviews are not biasing results. The incidence for a topic can then be described as

$$\text{Incidence(company, topic)} = \frac{\text{Number of reviews that mention topic}}{\text{Total number of reviews for a given company}}$$

We therefore have an incidence level of zero in the case in which no employees reviews opt to talk about that particular topic, and one in the case that all employees opt to talk about a particular value in their review. The summary statistics at the firm level are shown in Table 4.3

A large hurdle of dealing with text classifications is dealing with sentiment information. Our goals invovle correctly classifying whether a specific employee was talking about a certain culture trait, and whether that trait was spoken about in a positive or negative sense. Thankfully, by the nature of the prompts given to the user, the textual responses

Variable	Mean (%)	Std. Dev.	Q1	Q3
Agility	14.6	6.3	9.7	19.3
Collaboration	9.9	3.4	7.6	11.7
Customer Orientation	6.4	4.3	3.2	8.9
Engagement	88.3	2.9	86.5	90.4
Execution	11.1	4.6	7.8	13.9
Inclusivity	3.3	1.2	2.5	4.0
Innovation	7.2	5.1	3.2	10.3
Integrity	8.7	2.7	6.8	10.0
Performance	5.5	2.5	3.8	7.1
Respect	6.2	2.3	4.6	7.4

Table 4.3: This table shows the incidence levels of the different culture values as outputted from the classification model, aggregated at the company level.

are already sorted cleanly into their positive or negative sentiments. The variable *Pros* is overwhelmingly (99%) positive, *Cons* is overwhelmingly (99%) negative, and *Feedback* is overwhelmingly (98%) neutral or negative. This was confirmed both by manually checking a random sampling of thousands or reviews, and by running Stanford's *StanfordNLP* Python package created by their Natural Language Processing research group.

We can define the sentiment of a particular topic as some value between 0 and 1, representing the proportion of reviews that are deemed positive overall. The purpose of defining sentiment in this way allows us to treat each review equally, independent of their lengths or strength. For example, a single employee who leaves a lengthy and incredibly positive review should not outweigh multiple employees who leave concise, overall negative reviews.

Specifically, we define the net score of a particular review as

$$\begin{aligned}\text{NS(review, topic)} = &\sum_{term \in Value} \text{NrAppearances(review, term, pros)} \\ &- \sum_{term \in Value} \text{NrAppearances(review, term, cons)} \\ &- \sum_{term \in Value} \text{NrAppearances(review, term, feedback)}\end{aligned}$$

Any positive value of Net Score indicates an overall positive review for a particular value, and is consequently converted to the value 1. Alternatively, any net score value less than 1 is coded as 0. We now define the overall sentiment of a particular value for a given company as

$$\text{Sentiment(company, topic)} = \frac{\text{Number of Positive Reviews for Topic}}{\text{Number of Reviews that Mention Topic}}$$

In this formulation, an firm sentiment value of 0 indicates that all reviews about this topic were either completely neutral or negative, and a value of 1 indicates all reviews about this topic were positive. A benefit of maintaining sentiment as a positive-only value is the ability to apply distribution transformations with ease. The resulting classifications sentiment strengths are presented in Table 4.4

4.4 Analysis of Classification Output

As an overall assessment of these classifications, we find the correlations of these values against the one-to-five ratings on Glassdoor, as well as amongst themselves. Particularly, we would ideally expect moderately high, positive correlation between the sentiment values and Glassdoor defined metrics, and ideally little correlation between the values themselves

Variable	Mean (%)	Std. Dev.	Q1	Q3
Agility	13.3	6.3	9.2	16.7
Collaboration	37.6	9.5	31.2	43.4
Customer Orientation	38.3	15.3	27.7	46.7
Engagement	52.3	9.7	44.7	59.1
Execution	18.2	7.2	13.2	21.7
Inclusivity	19.6	10.7	11.9	25.0
Innovation	43.3	14.8	32.4	54.7
Integrity	20.1	11.6	11.7	26.1
Performance	19.3	9.1	13.2	23.6
Respect	15.3	10.4	8.4	19.0

Table 4.4: This table shows the sentiment levels of the different culture values as outputted from the classification model, aggregated at the company level.

to avoid multicollinearity issues. This follows from the notion that an employee who leaves a high rating on the one-to-five scale will be more likely to write an overall positive rating on the company as well, with variability in what they write their response about. We plot the correlation matrix in Figure 4.1. We see distinct groupings in how the correlation plot appears. For a certain classification value, it maintains fairly consistent and positive correlations with Glassdoor one-to-five ratings, but low and inconsistent correlations with other value classifications. Some notable exceptions are *Agility*, which maintains close to zero correlations across all Glassdoor metrics, and to a smaller extent *Inclusivity* and *Innovation*.

We start with the simplest model and work our way up. Specifically, we work to create a prediction model to predict four-factor alphas so see which predictors might be indicative of future alphas. We begin by assessing some underlying assumptions about our new data. We start by checking normality within the data by constructing both histograms and QQ-plots of the data, as seen in Figures 4.2 and 4.3

Looking at both the histograms and QQ-plots of the data, it appears that most of the

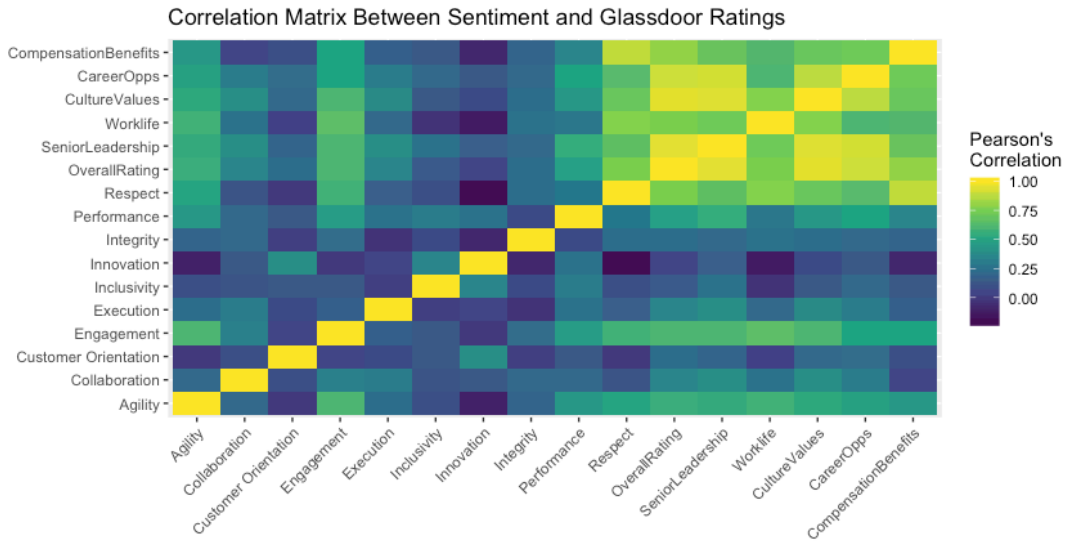

Figure 4.1: This heatmap shows the correlation between Classification sentiment measures and Glassdoor's one-to-five metrics. In particular, we see noticeably three distinct clusters emerge. Classifications are consistently positively correlated with the defined Glassdoor metrics, and not particularly correlated with each other.

Figure 4.2: In this figure we show the resulting distribution of the average sentiment scores by classification value, aggregated at the company level on July 1, 2018.

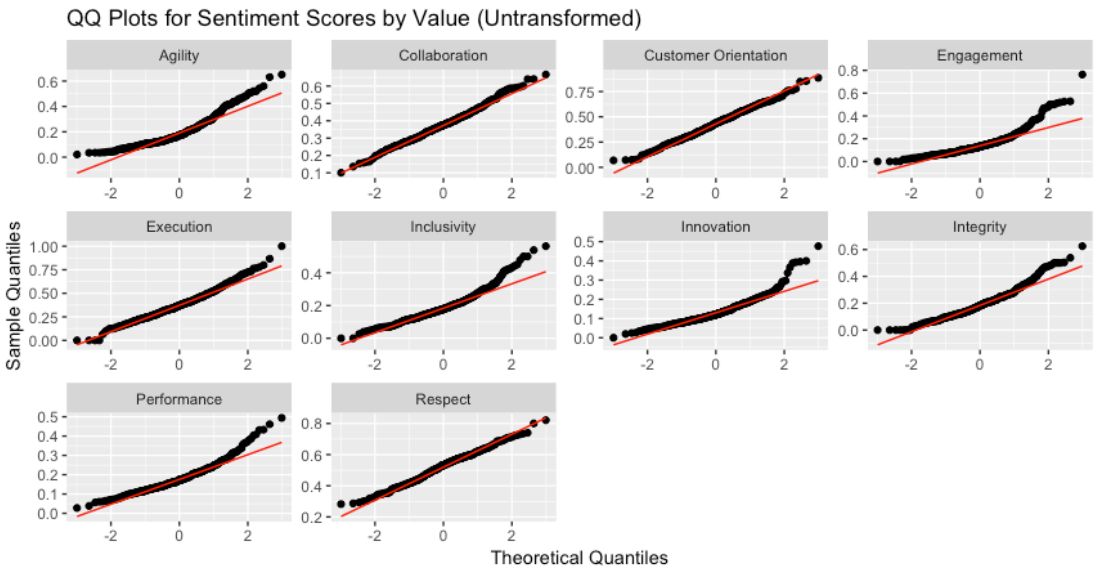

Figure 4.3: In this figure we show the resulting QQ distributions of the average sentiment scores by classification value, aggregated at the company level on July 1, 2018.

data appears approximately normal, though some do appear to have a bit of a right skew. Specifically, "Agility", "Inclusivity", "Engagement" and "Performance" are corrected through a square-root transformation. The square root transformation has the benefit over the log transformation due to working on zero values, and fixes the distribution without overcorrecting. While it may appear that Engagement post-transformation is still slightly right skewed, we opt to maintain only one square root transformation for interpretability benefits.

4.5 Panel Analysis Using Classification Output

We now look to create a model that could potentially explain a relationship between these classified reviews and a firm's valuation on the market. Similar to Chapter 2, we turn to performing Panel Data Analysis as each firm experiences its own time series of returns within

Value	Coefficient	t-stat	P-value
Agility	2.38	2.79	0.005
Collaboration	-0.93	-1.50	0.133
Customer Orientation	0.57	1.98	0.048
Engagement	4.33	3.78	0.000
Execution	0.20	0.22	0.827
Inclusivity	-0.07	-0.18	0.860
Innovation	-0.23	-0.48	0.626
Integrity	0.20	0.21	0.831
Performance	0.22	0.53	0.594
Respect	-0.38	-0.32	0.747

Table 4.5: This table reports the results for the Fama MacBeth regression when all values are included as predictors.

the broader economy. In particular, we again perform a Fama Macbeth cross-sectional regression on the monthly returns for each firm. Specifically, we have a model of the form

$$R_{i,t+1} = \gamma_{0,t} + \gamma_{1,t}\text{Value}_{i,t} + \gamma_{i,t}\text{X}_{i,t} + \varepsilon_{i,t}$$

where $\text{Value}_{i,t}$ represents the vector of values for firm i at month t by sentiment, $\text{X}_{i,t}$ represents the list of controls for firm i at month t included in the regression, and R_{i,t_1} represents the excess return for firm i in month $t+1$. The resulting significance values when regressed in the same model are shown in Table 4.5.

We see that when all values are put in the model, some values that we may naturally associate more strongly with firm returns are deemed significant. Specifically, *Engagement*, *Agility*, and *Customer Orientation* are values directly tied to the general effectiveness of a company and are significant at the $p = 0.05$ level. The positive coefficients for these three values is also a good sign, as traditional managerial economics would expect higher returns to be a result of a company's ability to serve the customer or enact new changes [citation].

Value	Coefficient	t-stat	P-value
Agility	1.59	2.41	0.016
Collaboration	-0.22	-0.73	0.465
Customer Orientation	0.74	4.72	0.000
Engagement	3.73	3.55	0.000
Execution	1.03	1.91	0.056
Inclusivity	0.04	0.12	0.908
Innovation	0.24	0.55	0.586
Integrity	0.99	1.53	0.126
Performance	0.84	1.60	0.109
Respect	0.10	1.22	0.226

Table 4.6: In this figure we present the results of performing the Fama Macbeth Regression when only one value is used as the predictor variable, while still including controls. For example, the row for "Agility" reflects the results when only Agility sentiment values and controls are used in the regression.

With the exception of *Collaboration*, which has an associated p-value of 0.133, the other variables in the model do not appear to be close to significant. The negative coefficients for some of the values are a bit unusual, but may be explained as a side effect of being included in the model while being insignificant. Finding the Fama MacBeth regression using only one value results in positive coefficients with the exception of *Collaboration*, yet results in no changes in the significance of variables. (Table 4.6)

4.6 Portfolio Sorting Using Classification Output

We now look to add validity to the regression by performing a similar portfolio sorting procedure described in Chapter 2. Similar to Chapter 2, we create three separate portfolios representing the lowest quintile, middle three quintiles, and highest quintiles managed at each quarter. We start first by performing portfolio sorting on the sentiment of each of the values at a given time. For example, $P_{H,t,Respect}$ represents a portfolio the highest 20% of firms as ranked by Respect's sentiment at quarter Q_t, and is held until quarter Q_{t+1}. We calcu-

late alpha and return values at some quarter Q_t as the based on the portfolio created at Q_{t-1} and held until Q_t. This portfolio sorting approach is performed for the values *Engagement*, *Agility*, and *Customer Orientation*. Because the incidence rates for the values are relatively low at the onset, we only consider firms that have at minimum 20 reviews classified for that topic at a given time. Significance of alphas described in the Cahart Four-Factor model are then computed using the Newey-West t-stat adjustment and shown in Table 4.7. We see that the High quintile portfolios generate significant alpha returns as well, though other portfolios generated are unable to reach significant vlaues.

	Portfolio	Return (%)	α (%)	t-stat	p-value
Agility	Low	0.68	-0.05	-0.43	0.666
	Med.	0.93	0.14	1.42	0.232
	High	1.45	0.65	4.05	0.000
Customer Orientation	Low	0.61	-0.08	-0.53	0.600
	Med.	0.95	0.11	1.54	0.574
	High	1.07	0.38	3.18	0.001
Engagement	Low	0.57	-0.02	-0.08	0.933
	Med.	0.94	0.15	1.32	0.188
	High	1.06	0.29	3.27	0.001
Overall Market		0.93			

Table 4.7: We show the results of portfolio sorting by the sentiment of each of the different metrics, individually. Alpha and Returns are given in a monthly format

4.7 Clustering Issues in High Performing Firms

We see in Table 4.7 that the portfolio sorting approach can also return some significant portfolios, similarly to results found in Table 3.1. However, we should note that many of the significant portfolios generated have similar significance and alpha values, particularly for those that are defined as *High* portfolios. This might indicate that the portfolios gener-

ated for the upper quintile are actually quite similar. Though the companies to this point haven't appeared to have serious multicollinearity issues, it might be that companies performing extraordinarily well in some value are also performing significantly well by sentiment in others. This turns out to be the case upon further examination, with 80% portfolio similarity between "Agility" and "Customer Orientation," 73% portfolio similarity between "Agility" and "Engagement," and 91% portfolio similarity between "Customer Orientation" and "Engagement."

4.8 Discussion

In this chapter we present results on performance based solely on the text reviews left by employees. Though the results depend largely on the underlying classification model, we are still able to witness significant portfolio selection based on certain values. Specifically, a company's "Agility," "Customer Orientation" and "Engagement" can all be used to select significant alpha portfolios, coninciding with research that relates positive customer orientation with significant growth over time[32]. "Agility" especially is often discussed as an important factor in determining the longterm success of a company's performance, and it's significance through text classification alone may suggest that employee's discussion about the topic is enough to determine near term performance[46].

5

Conclusion

5.1 Discussion

Though many of the results presented in this thesis are promising, it's important to discuss the context and limitations of the findings.

First, it is important to consider the time frame and general market trends in which all analysis was conducted. All excess returns and alpha coefficients are representation be-

tween the dates of January 1, 2014 and July 1, 2018, and therefore only represents 54 months of data. This time period is generally summarized as a period of strong growth within the market, and overall relatively short time frame in comparison to usual market analysis[33]. When comparing the results to peers such as Edmans, for example, we experience a marginally higher alpha relative to his computed across 25 years of data[19]. It's especially important to note that the time period we perform our analysis on does not include any official periods of recessions as defined by the National Bureau of Economic Research. It is therefore difficult to translate the results found in this period to future periods of distress in the market.

It's also important to consider the dataset used for the analysis. Specifically, the companies in question are all American and typically large companies, and the significance of portfolio returns was based on alpha significance against the American stock market. Alpha is most rigorously studied and modeled within the American economy, and therefore may result in uncharacteristic findings when translated across regions [citation]. Additionally, portfolio creation though a mixture of domestic and international stocks may result in ambiguous alpha terms as well[41].

Furthermore, the results presented in Chapter 4 on textual classifications are highly dependent on the quality of the underlying classification model. The keyword approach has the benefit of being able to specify exact words or phrases indicative of a single or multiple classifications, but requires a significant amount of manual oversight. Even with manual oversight, whether or not a text instance is truly describing a particular value can be a bit blurred and open to interpretation. Though every precaution was taken to reduce the number of errors, particularly in the number of false positives, it would be foolish to say the

model is perfect. The classification model should continuously be updated to reflect new findings and writing styles, but in its current state still presents promising findings.

5.2 Further Analysis

Much of the analysis presented in this paper related performance to alpha or excess returns, though there are many other measures of firm and portfolio performance that may achieve different nuanced conclusions. Other common portfolio performance metrics include the Sharpe Ratio or Treynor's Measure, which both similarly attempt to evaluate portfolio performance given its risk over time[43]. We can also take a more granular approach and relate the ratings or classifications to firm specific characteristics like changes in Return on Sales, EV\EBITDA, Tobin's Q, or various other firm specific ratios[38].

Additionally, the portfolios created in this thesis were solely interested in the long position. The negative alphas on some of the portfolios may be indicative that portfolio selection for the purpose of shorting stocks could also be a viable option and should be more rigorously explored[26]. The combination of both long and short positions may work to reduce portfolio variance and result in higher performance[16].

5.3 Conclusion

With this thesis we present results that add to the growing literature of research in this area. While much research to date has been focused on leadership and executives, we show that the aggregate views of employees may also be indicative of firm performance. Employees may generally have knowledge and insight about the firm's operations before the market fully recognizes, potentially reflected in their reviews of their employer via sites like Glass-

door. We match similar results by Edmans that the market generally undervalues the aggregated rankings by companies[19,20]. We also show that employees may be able to predict positive changes in senior management, supporting claims by authors indicating senior management's role in company success[12]. Lastly, we also sought to find more granular explanations that may exist in employee written reviews. Specifically, we show that the positive sentiment of certain topics including "Agility," "Customer Orientation," and "Engagement" may be indicative of market performance, matching claims by authors suggesting these values are key to long term growth[32,46].

A

Appendix

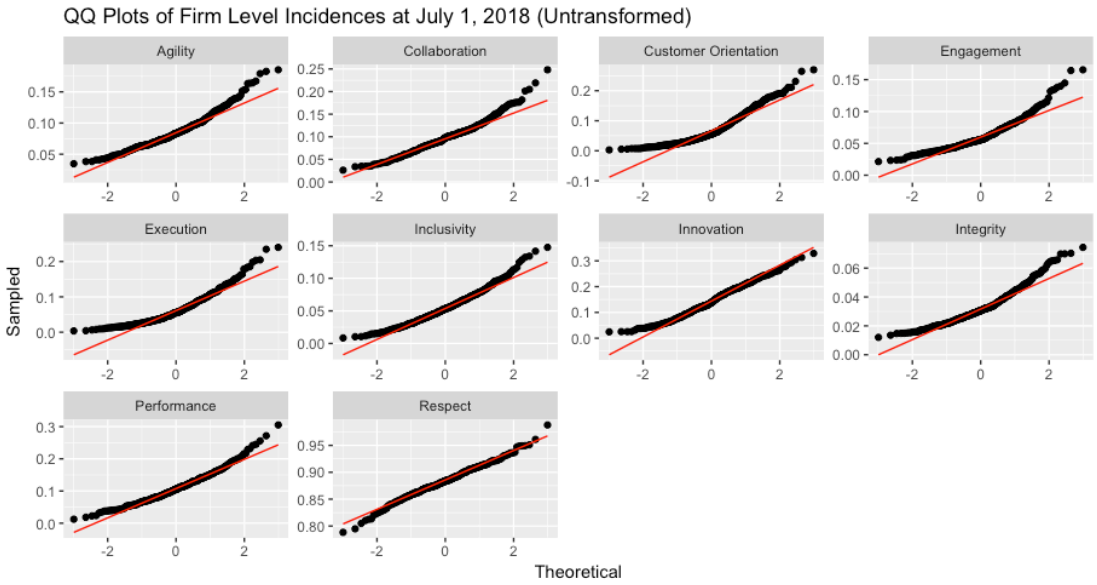

Figure A.1: This figure presents the resulting QQ plots for Incidence rates before applying any transformations

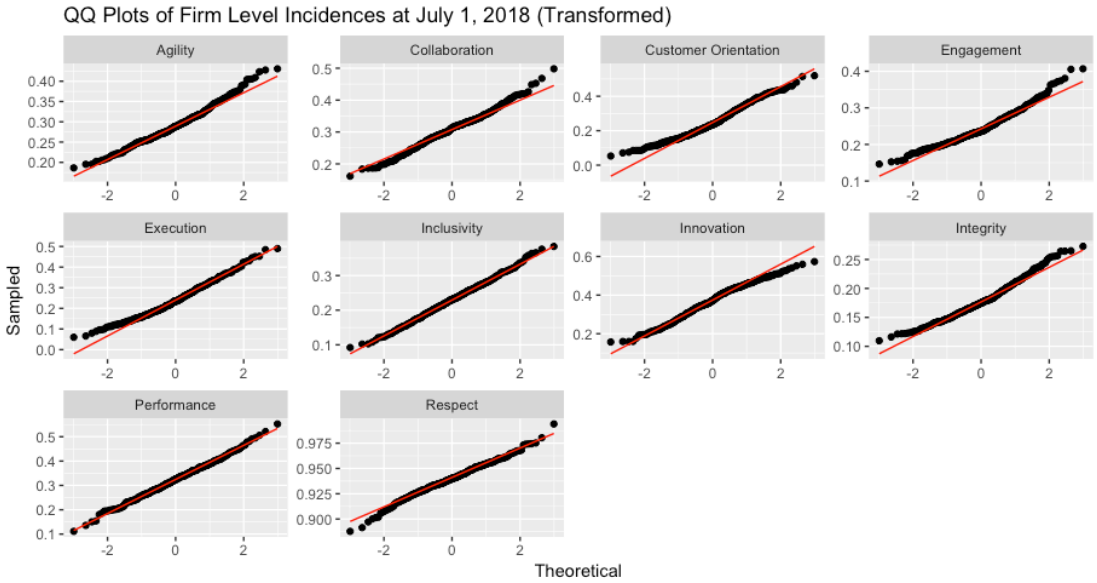

Figure A.2: This figure presents the resulting QQ plots for Incidence rates after applying a square root transformation.

The code and data use to produce the analysis is available on Github at https://github.com/PeterAAyala/Thesis_Git.

- The Fama MacBeth Regressions were performed using the "linearmodels" python package, with documentation available at https://bashtage.github.io/linearmodels/doc/index.html.
- The "sandwich" package was used initially as well for robust covariance matrics and Newey West T-statitstics, with documentation available at https://cran.r-project.org/web/packages/sandwich/sandwich.pdf
- Visualizations in this thesis were made using "ggplot2" in R, with documentation available at https://www.rdocumentation.org/packages/ggplot2/versions/3.1.0
- Due to an NDA signed with Glassdoor, I am unable to share the raw dataset containing the indiviual employee reviews, their resulting classifications on an individual level, or the code to produce the classification model itself, though sentiment and incidence levels by month and company to perform analysis are given.

References

[1] (2009). *Developing a corporate ethics strategy : leading CEOs on building a culture of trust, addressing ethical dilemmas, and ensuring company consistency*. Inside the minds. Boston]: Aspatore.

[2] Adalsteinsson, G. & Grimsdottir, E. (2015). The importance of company culture in a merged company. *International Journal of Business Research*, 15(3), 105–114.

[3] Arosa, C. M. V., Richie, N., & Schuhmann, P. (2015). The impact of culture on market timing in capital structure choices. *Research in International Business and Finance*, 35(1), 180–196.

[4] Babenko, I. & Sen, R. (2015). Do nonexecutive employees have valuable information? evidence from employee stock purchase plans. *Management Science*, 62(1), 1878–1898.

[5] Bau, F. & Wagner, K. (2015). Measuring corporate entrepreneurship culture. *International Journal of Entrepreneurship and Small Business*, 25(2), 231–244.

[6] Benmelech, E. & Frydman, C. (2015). Military ceos. *Journal of Financial Economics*, 117(1), 43–59.

[7] Biggerstaff, L., Ciceo, D., & Puckett, A. (2015). Suspect ceos, unethical culture, and corporate misbehavior. *Journal of Financial Economics*, 117(1), 98–121.

[8] Black, F., Jensen, M., & Scholes, M. (1972). The capital asset pricing model: Some empirical tests. *Studies in the Theory of Capital Markets*, (pp. 79–121).

[9] Blei, D., Ng, A., & Jordan, M. (2002). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

[10] Cahart, M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1), 57–82.

[11] Cegarra-Navarro, J.-G., Reverte, C., Gómez-Melero, E., & Wensley, A. (2016). Linking social and economic responsibilities with financial performance: The role of innovation. *European Management Journal*, 34(5), 530–539.

[12] Clarke, N. & Mahadi, N. (2017). Mutual recognition respect between leaders and followers: Its relationshop to follower job performance and well-being. *Journal of Business Ethics*, 141(1), 163–178.

[13] Cohen, L., Malloy, C., & Pomorski, L. (2012). Decoding inside information. *Journal of Finance*, 67(3), 1009–1043.

[14] Conway, N. & Briner, R. (2002). Full-time versus part-time employees: Understanding the links between work status, the psychological contract, and attitudes. *Journal of Vocational Behavior*, 61(2), 279–301.

[15] Davidson, R., Dey, A., & Smith, A. (2015). Executives' "off-the-job" behavior, corporate culture, and financial reporting risk. *Journal of Financial Economics*, 117(1), 5–28.

[16] Davis, P. (2017). An optimal mix of factors. *Financial Analysts Journal*, 73(4), 37–39.

[17] Duan, W., Gu, B., & Whinston, A. (2008). Do online reviews matter? an empiritcal investigation of panel data. *Decision Support Systems*, 45(4), 1007–1016.

[18] Eccles, R., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. *National Bureau of Economic Research*, 60(11), 2835–2857.

[19] Edmans, A. (2011). Does the stock market fully value intangibles? employee satisfaction and equity prices. *Journal of Financial Economics*, 101(3), 621–640.

[20] Edmans, A., Li, L., & Zhang, C. (2014). Employee satisfaction, labor market flexibility, and stock returns around the world. *NBER Working Paper Series*.

[21] Ernstberger, J., Haupt, H., & Vogler, O. (2011). The role of sorting portfolios in asset-pricing models. *Journal of Machine Learning Research*, 21(18), 1381–1396.

[22] Evbayiro-Osagie, E. I. & Osamwonyi, I. O. (2017). A comparative analysis of four-factor model and three-factor model in the nigerian stock market. *International Journal of Financial Research*, 8(4).

[23] Fama, E. & French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2), 427–465.

[24] Fama, E. & French, K. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.

[25] Ferrell, A., Liang, H., & Renneboog, L. (2016). Socially responsible firms. *Journal of Financial Economics*, 122(3), 585–606.

[26] Fung, W. & Hsieh, D. A. (2011). The risk in hedge fund strategies: Theory and evidence from long/short equity hedge funds. *Journal of Empirical Finance*, 18(4), 547–569.

[27] Galema, R., Plantinga, A., & Scholtens, B. (2008). The stocks at stake: Return and risk in socially responsible investment. *Journal of Banking Finance*, 32(1), 2646–2654.

[28] Guiso, L., Sapienza, P., & Zingales, L. (2015). The value of corporate culture. *Journal of Financial Economics*, 117(1), 60–76.

[29] Hart, C. E., Lence, S. H., Hayes, D. J., & Jin, N. (2016). Price mean reversion, seasonality, and options markets. *American Journal of Agricultural Economics*, 98(3), 707–725.

[30] Hildreth, J. A. & Anderson, C. (2016). Failure at the top: How power undermines collaborative performance. *Journal of Personality and Social Psychology*, 110(2), 261–286.

[31] Hogan, S. & Coote, L. (2014). Organizational culture, innovation, and performance: A test of schein's model. *Journal of Business Research*, 67(8), 1609–1621.

[32] Homburg, C., Muller, M., & Klarmann, M. (2011). When does salespeople's customer orientation lead to customer loyalty? the differential effects of relational and functional customer orientation. *Journal of the Academy of Marketing Science*, 39(6).

[33] Issler, J. V. & Vahid, F. (2006). The missing link: using the nber recession indicator to construct coincident and leading indices of economic activity. *Journal of Econometrics*, 132(1), 281–303.

[34] Jiu, M., Wolf, C., Garcia, C., & Baskurt, A. (2012). Supervised learning and codebook optimization for bag-of-words models. *Cognitive Computation*, 4(4), 409–419.

[35] Kim, K. & Watkins, K. (2017). The impact of a learning organization on performance. *European Journal of Training and Development*, 41(2), 177–193.

[36] Kratzer, J., Meissner, D., & Roud, V. (2017). Open innovation and company culture: Internal openness makes the difference. *Technological Forecasting and Social Change*, 119(1), 128–138.

[37] Krüger, P. (2015). Corporate goodness and shareholder wealth. *Journal of Financial Economics*, 115(2), 304–329.

[38] Laura, V. (2011). Assessing the firm performance through the financial ratios. *Analele Universităţii Constantin Brâncuşi din Târgu Jiu : Seria Economie*, 1(3), 159–166.

[39] Lerner, J. & Tirole, J. (2002). Some simple economics of open source. *Journal of Industrial Economics*, 50(1), 197–234.

[40] Liu, X. (2016). Corruption culture and corporate misconduct. *Journal of Financial Economics*, 122(2), 307–327.

[41] Maccheroni, F., Marinacci, M., & Ruffino, D. (2013). Alpha as ambiguity: Robust mean-variance portfolio analysis. *Econometrica*, 81(3), 1075–1113.

[42] Newey, W. & West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(1), 703–708.

[43] Nielsen, L. T. & Vassalou, M. (2004). Sharpe ratios and alphas in continuous time. 39(1), 103–114.

[44] O' Reilly III, C., Caldwell, D., Chatman, J., & Doerr, B. (2014). The promise and problems of organizational culture: Ceo personality, culture, and firm performance. *Group Organization Management*, 39(6), 595–625.

[45] Ohunakin, F., Adeniji, A., Oludayo, O., & Osibanjo, O. (2018). Perception of frontline employees towards career growth opportunities: implications on turnover intention. *Business: Theory and Practice*, 19(1), 278–287.

[46] Phillips, P. (2004). How agile is your company? agility is an necessity in today's fast-paced and competitive marketplace. *Coatings World*, 9(8).

[47] Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, 56(1), 18–36.

[48] Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10).

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (XII) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.