香 港 大 學
THE UNIVERSITY OF HONG KONG

Half Full or Half Empty?

A Sentiment Analysis of Glassdoor Textual Ratings Data and Opportunities in Quantitative Trading

Jadon Ng Tsz Hei, 3036076067
Rhenald Louwos, 3035756751
Poon Tsz Chung, 3036077774
Theo Obadiah Teguh, 3035898872

FINA4359 Data Analytics, Quantitative Finance, and Blockchain Finance
Group Research Report

HKU BUSINESS SCHOOL
港大經管學院

---

1. Introduction

1.1 Advances in NLP

In recent years, the field of Natural Language Processing (NLP) has grown exponentially due to methodological advancements and increasing data availability (Xing et al., 2018). Feature engineering with word embedding technologies such as Word2Vec (Li et al., 2021; Chandola et al., 2023) have enabled researchers to represent text data as numerical vectors which then serve as training data for machine learning models (Osterrieder, 2023). Building upon these vector representations, text mining technologies such as sentiment analysis or opinion mining have become one of the tools that are predominantly utilized in numerous sectors (Gupta et al., 2020). In particular, advanced machine learning-based frameworks approach sentiment analysis as a classification task performed on textual datasets. In light of these advancements, there have been various promising attempts to use this sentiment analysis framework for statistical forecasting in economics and finance, which led to the establishment of the natural language-based financial forecasting (NLFF) research field (Xing et al., 2017). For instance, Muthukumar and Zhong (2019) proposed a novel stochastic time series generative adversarial network built on Naive Bayes' sentiment analysis synthesizing financial data and text for stock price forecasting. These applications have also led to the development of NLP tools to generate trading signals and facilitate risk management in finance (Osterrieder, 2023).

Despite the aforementioned facts, machine learning-based sentiment analysis still faces major challenges, as these approaches require domain-specific datasets (Gupta et al., 2020). Moreover, the high economic cost of creating intensive machine learning systems from the ground up has led to the initiation of open-source platforms such as HuggingFace, a popular online hub and repository for pre-trained machine learning models (Pepe et al., 2024). These platforms allow public access to state-of-the-art pre-trained models such as Google's BERT and Facebook's RoBERTa which streamlines the process of vectorization and sentiment analysis.

1.2 Navigating Glassdoor

In this research project, our team analyzes a Glassdoor data set comprising 3.4 million textual reviews and numerical ratings on US-listed firms across various industries. As Glassdoor is a US-based website where current and former employees anonymously review companies, we aim to extract trading signals utilizing the machine learning-based sentiment analysis framework on this domain-specific database. Firstly, we will explain the foundational aspects of our dataset tables and the findings of our exploratory data analysis processes. Afterward, we propose a BERT-based model, coupled with thematic sentiment analysis to mine alpha-predictive factors. Each of the generated factors is then integrated within specific mathematical transformations to generate signals in various trading strategies tested in our in-house self-implemented backtesting platform. In addition, we discuss the computing hardware requirements and model training time needed to reproduce our study. Future ideas and potential innovations will also be reviewed.

2. DataFrame Overview

This section introduces several aspects of our research database comprising three main data frames: *company*, *glassdoor_classification,*

textual_review, and an additional database of corresponding historical returns.

2.1 *company*

This table provides information about the companies in our universe. This includes the revenue level, company size, and industry category provided by Glassdoor. A total of 5018 companies are included.

We apply filters on the tickers by quarterly and total review count and later ensure they are traded in the US from 2020-2023. Different filters are created, including filtering on a minimum number of reviews per month or quarter, filtering on minimum market capitalization, etc. Depending on the nature of the trading strategy, we will apply different filters to the stock universe as a preliminary universe selection for noise reduction, which are specified explicitly in the latter sections.

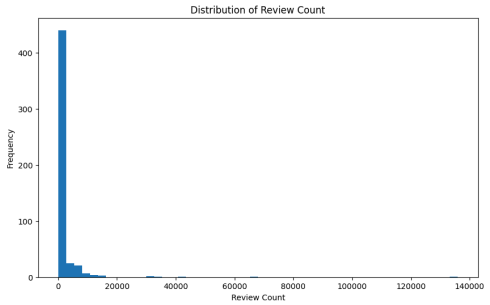

Fig 2.1. Review Count Distribution
(Note that most companies have limited reviews)

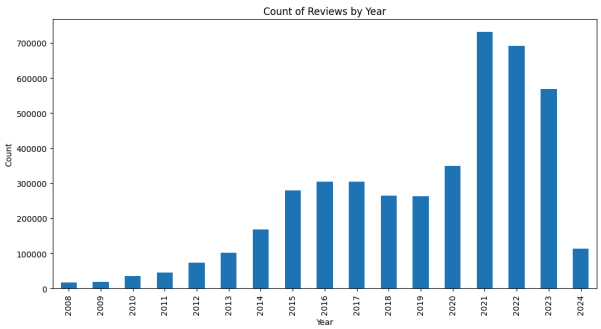

Fig 2.2. Review Count By Year

2.2 *glassdoor_classification*

This data frame is the backbone of our analysis, bringing 4,337,584 reviews from 2008 to 2024. More descriptions and a comprehensive exploratory data analysis of the Glassdoor rating data frame will be included in later sections.

Here, numerical ratings are defined on an ordinal scale of 1-5 on different aspects given by users. These include an *Overall rating, Culture and Values, Senior Leadership, Work-Life Balance, Career Opportunities, Diversity, and Inclusion.*

Categorical ratings such as *Business Outlook* (Positive, Neutral, Negative), *Ceo Approval rating* (Approve, No Opinion, Disapprove), and *Recommend to Friends* (Positive, Negative) are also included.

The data also provides the opinion of other users on a specific review being *Helpful* or *Not Helpful*, representing the number of people who agree with the relevancy of the review.

2.3 *textual_review*

For each comment on Glassdoor, the reviewer will put a verbal comment on *Summary, Pros, Cons,* and *Advice*, which has a unique identifier that can be mapped or traced back to the *glassdoor_classification* data frame. Below are some comment samples.

"Stable pay and comprehensive benefits for family"
"No commission, so that could be hard to stay motivated and goal-oriented. Hard to get promotions because of the sheer size of the teams"

These may provide a more comprehensive understanding of the employee's perspective from a pure Likert scale. More description and

exploratory data analysis on the textual data will also be included in later sections.

3. Numerical Ratings Data

3.1 Exploratory Data Analysis

We first conduct the study by analyzing the numerical rating data from the Glassdoor reviews. Particularly, we first construct the monthly aggregated data frame of the key categories defined in Glassdoor: *Overall rating, Compensation and Benefits, Senior Leadership, Work-Life Balance, Culture and Values,* as well as *Career Opportunities*. Note that since the *ratingOverall* is required for all Glassdoor reviews, we first drop all the reviews with *ratingOverall* missing. Below is a table summary of the missing values of the data frame.

Table 3.1. Missing values for each rating

rating	missing (%)
Overall	0
Comp. & Benefits	800743 (19.0%)
Senior Leadership	843583 (20.0%)
Work-Life Balance	810539 (19.2%)
Culture & Values	959833 (22.7%)
Career Opportunities	791133 (18.7%)

We preprocess the data by aggregating each review into a monthly representation with operations applied to each column individually, meaning that missing values will be dropped if a specific rating is missing for each review. Then, we also access the monthly aggregated distributions of the six numerical ratings. Specifically, we visualize the 12-month rolling average aggregated distribution for smoothening purposes.

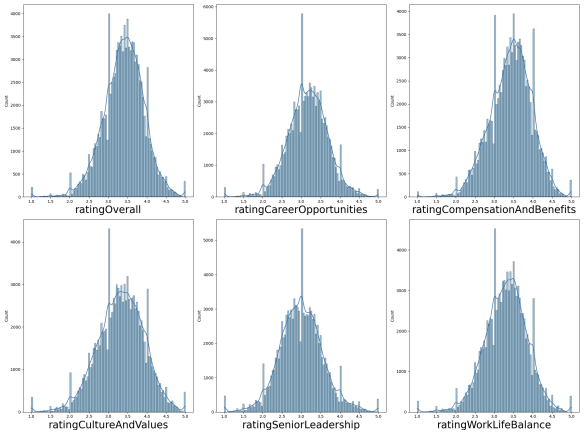

Fig 3.1. Distribution of monthly aggregated ratings

From the distributions, we can observe that the ratings follow a Gaussian-like distribution with the mean lying slightly above 3 for all of the ratings except for *SeniorLeadership*, which has a mean slightly below 3. The QQ plot for the above distribution can be found in the appendix (fig A.1.).

Apart from the distribution, we also visualize the correlation matrix between the numerical ratings, which exhibits high cross-correlation as expected, with an average correlation of around 0.75. This makes it challenging to find orthogonal signals with just numerical rating data alone, and it is a major reason for us to explore textual analysis on the given textual review in later sections.

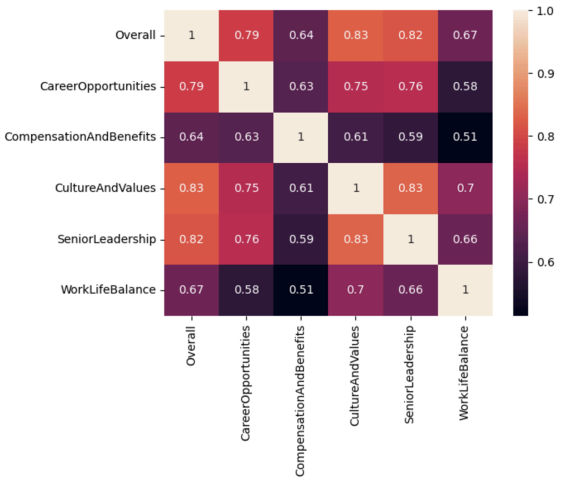

Fig 3.2. Correlation matrix of numerical ratings

To understand how the average ratings move over time, we also visualize the monthly average ratings from 2014-2022, noting that this plot will come into play in a later subsection.

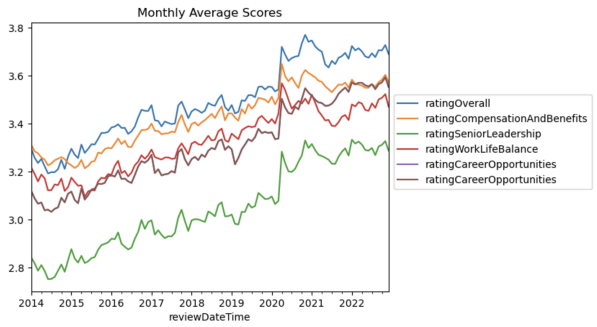

Fig 3.3. Monthly average scores 2014-2022

Interestingly, we see the average ratings demonstrating a non-stationary characteristic, a consistent uptrend with the ratings spiked during COVID-19 and a slight drawdown afterward.

3.2 Regression Analysis

Upon proper cleaning and aggregation of data, we first begin our analysis by conducting regression analysis to determine if there is a statistical relationship between numerical rating and returns. Specifically, by sorting the stock universe based on the monthly aggregated rating values and binning them into three different portfolios (bottom 20%, 20%-80%, top 20%), we conduct the CAPM regression analysis defined as,

$$E[R_i] = \alpha + \beta(E[R_m] - R_f) + \epsilon_i$$

with monthly portfolio return as the dependent variable and the risk-free discounted market return as the independent variable. By running the CAPM regression framework, we aim to access the Jensen's alpha to statistically determine if one portfolio is more capable of generating abnormal returns (returns not explained by market returns) than the other. The result of the regression results is shown below (detailed regression statistics can be found in the Appendix Table A.1.). Note that the analysis is conducted from 2014-2022 with the portfolios being rebalanced monthly (ie. we pick each portfolio every month based on sorting of ratings), filtering away stocks with less than 5 Glassdoor reviews per month for each rolling month of interest to reduce the noise in the dataset.

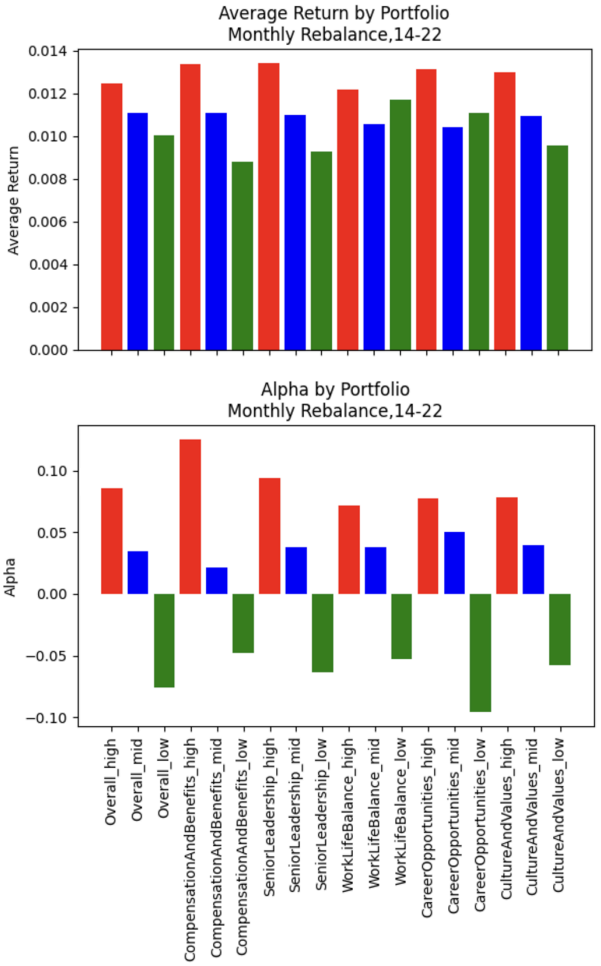

Fig 3.4. Regression result on numerical ratings

Through the bar charts, we can see that the monthly average returns are generally highest for high portfolios and lowest for low portfolios across the 6 numerical ratings. At the same time, it also demonstrates a relatively high alpha for high portfolios and a negative

alpha for low portfolios for all ratings other than *ratingWorkLifeBalance*.

3.3 Backtesting Results

While the regression framework allows us to conduct hypothesis testing to develop insight into the relationship between ratings and return, we have also run different trading signals to better analyze and evaluate the portfolio return over time. For instance, below shows the backtesting result of a monthly rebalanced, market-neutral portfolio, with the weight of each individual stock in the universe generated over on the signal:

rank(moving_average(numeric_rating, period))

Note that market neutral refers to the linear transformation of the weights of individual stocks such that the individual stock weights in the universe sum up to 0 and the absolute stock weights sum up to 1, making the portfolio more robust to market movements. We also take the risk-free rate to be 0.84%, which is the average US interest rate through 2014-2022 for convenience.

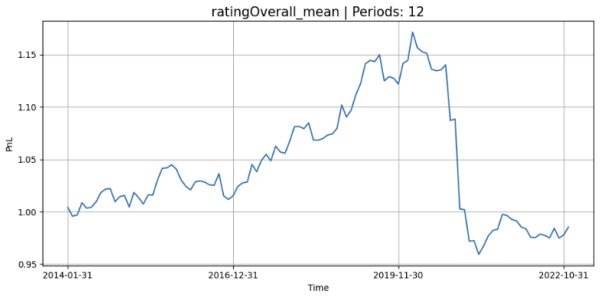

Fig 3.5. Backtesting results, 12-month rolling average on overall rating

3.4 Unemployed Reviews

In each Glassdoor review, the user may disclose whether they are currently employed. By decomposing the average ratings over time for both the unemployed and employed reviews, we can see that the upward trend of average ratings mentioned is largely caused by the employed review, while the average ratings for the unemployed group are more stable over the years..

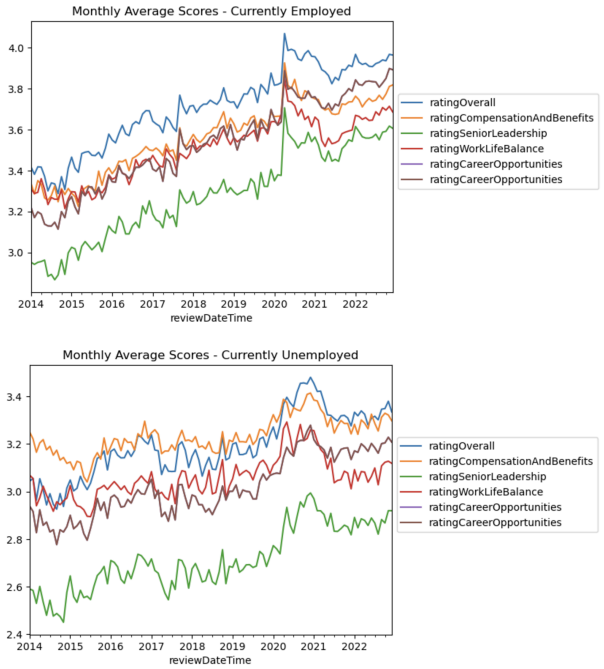

Fig 3.6. The monthly average score for employed-only and unemployed-only reviews

We also visualize the review count for both employed reviews and unemployed reviews in order to ensure that the data points are sufficient and are not significantly skewed to conduct our further analysis.

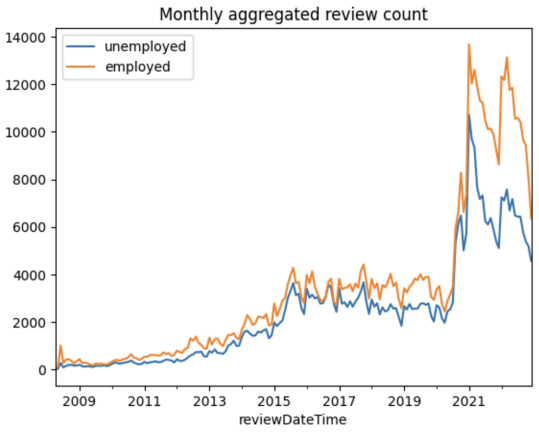

Fig 3.7. Monthly review count for unemployed and employed reviews

By identifying the key difference between the time series characteristic between the employed group and the unemployed group, we conducted further analysis by conducting the same regression (full regression result in appendix Table A.2.) and backtesting framework on the unemployed reviews only.

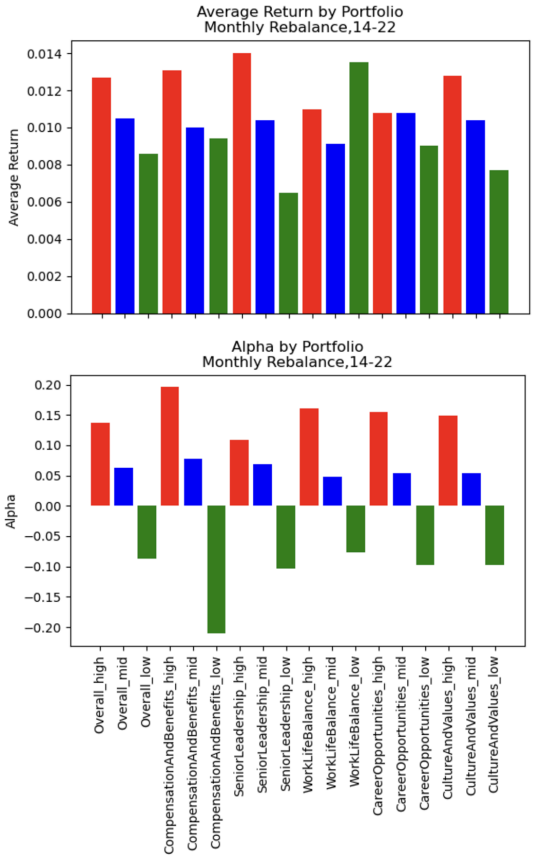

Fig 3.8. Regression results on unemployed reviews only

The regression result demonstrates similar characteristics to the unfiltered regression.

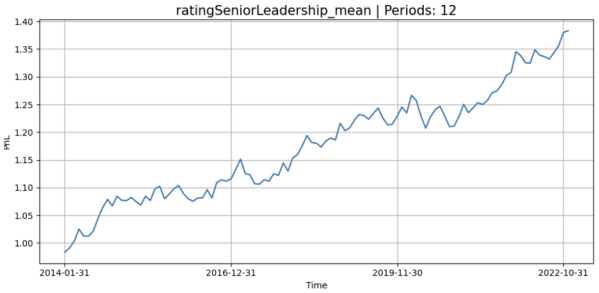

Fig 3.9. Backtesting results, SeniorLeadership on unemployed reviews

On the other hand, the backtesting results (with the same setup as the previously mentioned backtesting settings), particularly the backtesting result ran on *SeniorLeadership* demonstrated robustness through the COVID period, with a cumulative return of 38% and a Sharpe ratio of 0.7, distinguishing itself from other rating metrics which shows a significant drawdown from 2020 onwards, as well as demonstrating a significant difference when we ran the backtest only on the unemployed reviews rather than running it on all reviews together, in which we found to be very interesting coupled with the previous findings in the difference between unemployed and employed reviews.

4. Textual Review Data

4.1 Exploratory Data Analysis

In addition to the numerical ratings provided by Glassdoor, the textual reviews offer valuable insights that may help explain historical stock performance. After applying our filtering criteria, we obtained a total of 3,431,066 employee reviews. These reviews are comments or messages submitted by employees about their experiences at various companies on the Glassdoor platform.

summary	pros	cons	advice
Great benefits! managed very poorly	Great benefits, including 401k and HSA.	No training,& rude co-workers it\'s all about the color of shirt you have there..	I think this company should take matters seriously and not retaliate.make tour workers happier..simple little things like a little music, more incentives,and definitely update the IT department..quit terminating good employees that have spent many many years there, or even months.

Fig. 4.1. Textual section sample from the glassdoor dataset

4.1.1 Review Categories

The Glassdoor dataset consists of four distinct review categories, each representing different aspects of employee feedback.

1. *Summary*

As the name suggests, this section provides a brief, compact overview of the company at the time of the review. For our purposes, we interpret the *summary* as the primary message that the reviewer wants to convey about the company—essentially, their high-level opinion of the company's strengths or weaknesses.

2. *Pros*

This section contains comments on the positive aspects of the company. While it is generally expected to focus on the "good" features, it is important to note that not all comments in this category are entirely positive. Some reviews labeled as *pros* may include mixed sentiments or address issues that may not be wholly favorable.

3. *Cons*

Similarly, the *cons* section highlights the negative aspects of the company. However, much like the *pros* section, not all *cons* reviews are entirely negative as some include constructive criticism or neutral feedback, rather than outright negative remarks.

4. *Advice*

This section provides advice or suggestions from the employee, either for the company, prospective employees, or both. It is often rich in valuable insights, as seen in the sample above. Notably, the *advice* section ranks second in terms of average word count, reflecting its depth and detail compared to the other sections.

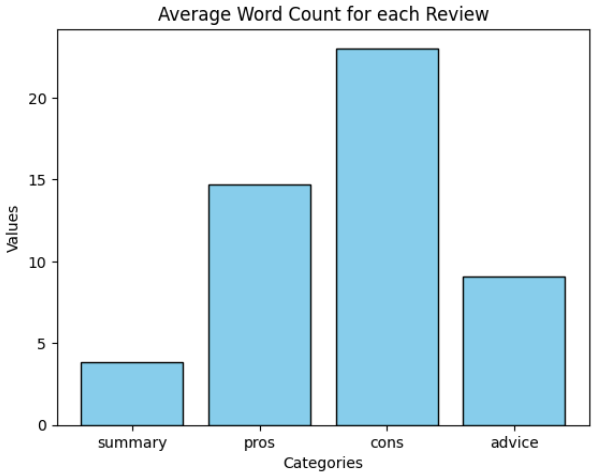

Fig 4.2. Comparison of average words per review across categories

4.1.2 Observations and Tone of Reviews

Upon analyzing the reviews, we observed that the language used in many of the responses, especially in the *cons* section, often carries a cynical or critical tone. This is particularly evident in the reviews with the highest word count, which tend to come from the *cons* section. The tone of these reviews is largely negative, reinforcing the general pattern of dissatisfaction expressed in the *cons* category.

summary	pros	cons	advice
Look elsewhere	Maybe pay and benefits only because if you need a job it's better than nothing. Managers do not want you here long term, they want to see turn over. I believe it's a way to remove higher salary folks and replace them with either contractors or lower skill folks so they can pay them less. I've seen people with 20 years with the company looked down on, like why are you still here. If you want a short term gig, I guess it's a great environment. Upper management treats people like widgets, it's easier to lay them off if you don't see them as people and view them as a number. They seem to have a higher regard for contractors than they do for employees, it's a really odd sight to see.	Anyone above a first level manager will not true you the truth so if they tell you something, expect the opposite. CITO rules IT with an iron fist. Anyone that does not support her is gone, everyone below her is just a yes person. Majority of managers are from retail, they have no clue how to run a utility based IT shop. They lied to everyone when they outsourced 60-75% of the departments with off shore contractors that have no right to call themselves an IT Engineer. Learn to support yourself because they can not support any of the systems, Google will become your best friend in fixing issues. My guess that's exactly what IT wants people to do, don't call in to talk to someone, fix it yourself.	Why bother, they won't listen. If you put up too much of a stink, they'll find a way to remove you. So keep your head down, do your job and hopefully the next CEO will see how big of a mess IT is and finally fire Therace and everyone that she brought in with her from retail. To the new CEO, fire the CITO on your first day, please for the sake of saving what little is left of IT and bring in someone from a utility background with no retail experience.

Fig 4.3. An example of the reviews in each category

This pervasive cynical tone has implications for our sentiment analysis model, which will

be further discussed in the NLP section of this paper. Given the skewed nature of the language, particularly in the *cons* reviews, we chose a sentiment model that is better suited to handle the more negative and sarcastic tone present in these reviews.

4.2 Approach and Reasoning

Our approach to utilizing the textual reviews dataset focuses on creating features that can influence a company's performance and, by extension, its stock price. As highlighted in the previous section, these reviews contain valuable insights into the internal dynamics of the company, which can directly or indirectly impact its market performance.

While each of the review categories contain different core messages, we combine them into a single column of reviews for several key reasons.

1. Feature Extraction Efficiency

By combining all reviews into one column, we can generate a larger pool of features in a single step. This aggregation enables a more comprehensive analysis of the data, as we can capture a broader spectrum of information from the same set of reviews.

2. Inconsistent *Pros* and *Cons*

The *Pros* and *Cons* categories do not always align with positive and negative sentiments, respectively. In some cases, a *pros* review may contain mixed sentiments or not be entirely positive, and similarly, a *cons* review may not be entirely negative. By consolidating the reviews, we eliminate the need to manually distinguish between positive and negative feedback and can instead analyze sentiment and performance on a more holistic level.

3. Avoiding Redundant Features

If the same topic or subject is mentioned across multiple review categories, we would risk duplicating feature extraction. Instead of recalculating features for each category separately, we combine them to ensure that each subject is analyzed just once, streamlining the feature extraction process. If duplication does occur, we would need to apply a weighted average to determine the final feature value. However, determining a uniform weight for this aggregation is challenging and may not always be straightforward after feature extraction.

4. Holistic Analysis of Topics

By merging the review categories, we can better understand how the different subjects or topics discussed in the reviews are interrelated. This collective approach enables us to capture the "bigger picture" of the company's performance, as we are not constrained by the artificial boundaries of individual review categories. This, in turn, allows us to more accurately assess the overall sentiment and health of the company based on the entirety of the feedback provided.

In summary, consolidating the review groups into a single column enhances the efficiency of our feature engineering process while enabling a more accurate and comprehensive analysis of the company’s performance indicators.

4.3 Methodology

In order to apply our approach to the textual dataset, we devised a two-step processing framework. The first step involves determining the topics or subjects stated in each review, while the second step determines

the value or sentiment of each topic identified from the reviews. The value of these topics is measured by the sentiment conveyed in the review text corresponding to each topic. Thus, our feature engineering pipeline integrates both topic modeling and sentiment analysis. We will first describe our sentiment analysis approach, followed by the topic modeling process.

4.3.1 Sentiment Analysis

Sentiment analysis is a natural language processing (NLP) technique used to determine the sentiment or emotion expressed in a piece of text. It classifies text into categories such as positive, negative, or neutral, allowing us to assess the general sentiment of the reviews towards various topics.

1. Model Input-Output

- **Input:** The textual reviews provided in the dataset.
- **Output:** A sentiment score for each review ranging from -1 to 1, which classifies each review as positive, negative, or neutral.

2. Model Selection

Our approach for sentiment analysis initially involved a trial-and-error method, where we experimented with different sentiment analysis models. We manually evaluated a random selection of 5-10 reviews to determine whether the sentiment scores were accurately aligned with human perception of the text. Although this was not an ideal method, it provided useful insights into the performance of various models.

To further refine our approach, we compared sentiment analysis on reviews to the sentiment of similar texts found on Twitter, as this would help ensure our models are generalizable across different platforms. Ultimately, we selected BERT-Twitter by Barbieri et al. (2020) for its accuracy in handling social media text, which tends to be more informal and contains slang and abbreviations that other models might miss.

4.3.2 Topic Modeling

Topic modeling is an unsupervised learning technique used to discover abstract topics from a collection of documents. We utilized Latent Dirichlet Allocation (LDA**)**, a popular topic modeling algorithm which assumes that each document is a mixture of topics, where each topic is a distribution over words. By analyzing these distributions, we can extract coherent topics from a set of reviews.

1. Overview of LDA

LDA is a probabilistic model used to uncover hidden thematic structures in a text corpus. Each topic is characterized by a distribution of words, and each document or review is represented as a mixture of these topics.

The model assigns topics to words and documents based on their co-occurrence patterns. During training, the model iteratively adjusts the topic assignments to maximize the likelihood of the observed words in the corpus.

In terms of computing load, LDA can be computationally expensive, especially with large datasets. Initially, we used a CPU-based approach for LDA, but performance was slow for large datasets (e.g., 10,000 reviews).

2. Optimization with CuPy

To enhance the speed of topic modeling, we incorporated CuPy, a GPU-accelerated library

that enables faster matrix operations. By using CuPy's vectorization capabilities before parsing the data into LDA, we observed significant improvements in performance. Specifically, we utilized Singular Value Decomposition (SVD) to preprocess the reviews, reducing dimensionality and making LDA processing more efficient.

Once CuPy was introduced, LDA performed considerably faster without sacrificing model accuracy. We benchmarked the runtime performance of our pipeline for various dataset sizes and found that LDA, when optimized with CuPy, significantly reduced the processing time of 10,000 reviews by 50%.

4.4 NLP Model Results

After applying sentiment analysis and topic modeling, we obtain a DataFrame containing the following features:

- Sentiment scores for each review (positive, negative, neutral).
- Topic distributions for each review, indicating the probability of each topic being present.
- Topic values, derived from the sentiment of the review associated with each topic.

The exact metrics used in the final DataFrame should aim to select the most meaningful features for subsequent analysis. In this paper, below are the list of metrics that we use as inputs for our NLP model:

- Employee Satisfaction
- Communication Effectiveness
- Workload Management
- Professional Development
- Work-Life Balance
- Team Collaboration
- Leadership Quality (standard of leadership)
- Innovation Encouragement
- Career Advancement Opportunities
- Employee Recognition
- Management Quality
- Benefits and Compensation
- Company Culture
- Job Security
- Productivity
- Quality of Work
- Credibility
- Leadership (general)

The DataFrame will then be used to generate alphas for our Portfolio Backtesting process.

4.5 Total Runtime and GPU Usage

We will now present the total runtime for the entire NLP pipeline, from sentiment analysis to topic modeling, along with GPU usage statistics to demonstrate the scalability of our approach.

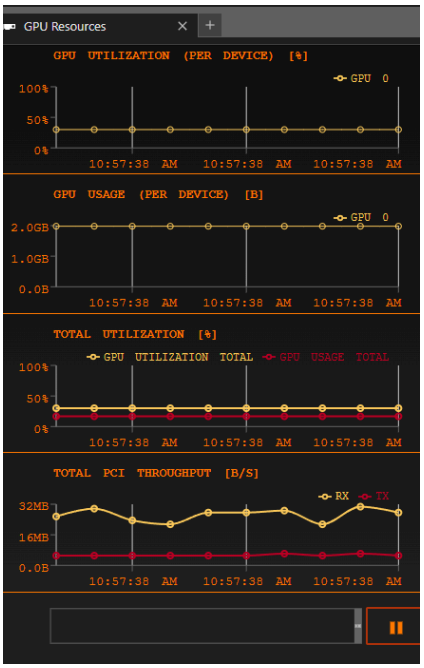

Fig 4.4. GPU utilization

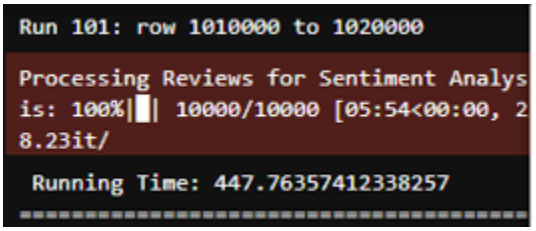

Fig 4.5. Average running time for 10,000 reviews

The above snippet is the average runtime for a batch of 10,000 reviews. The runtime ranges from 200-450 seconds, depending on the length of reviews on a single NVIDIA RTX 2080Ti that we obtained from the HKU School of Computing and Data Science GPU Farm through SSH. Due to the nature of SSH, we commonly encountered runtime issues which prevented a single offload of the complete dataset for the NLP process. As a result we resorted to batch runs. A batch saving process was made in place to store the feature results of each batch. The runtime can be further shortened through increasing the inner batch size on the code level. During some parts of the NLP notebook development we also utilized Google Colab extensively for their freely available GPU units. This made the complete NLP process fail-safe, robust, flexible, and scalable.

4.6 Backtesting Results

Upon aggregating the NLP processed dataframe, we first conduct some preliminary analysis to better understand the dataframe and distribution. We first visualize the distribution of the 12-month positive sentiment mean and the 12-month negative sentiment mean, noting that we take the 12-month average for better smoothing purposes, given how noisy the nature of the NLP generated data is (QQ plot, as well as positive or negative sentiment distribution in appendix).

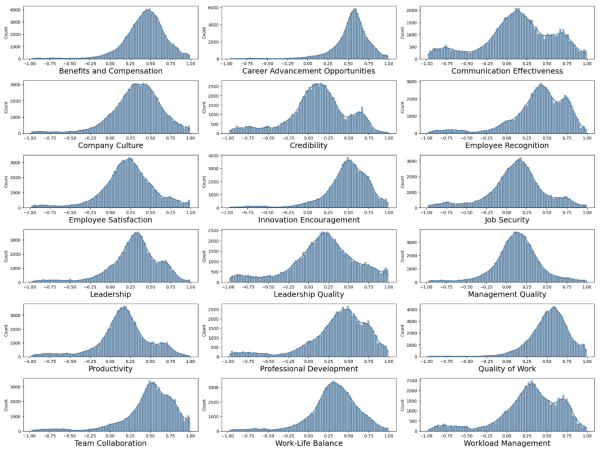

Fig 4.6. Distribution of 12-month sentiment mean

Upon visualizing the distribution, we also plot the correlation matrix to see how the NLP-generated features are correlated with each other.

Fig 4.7. Correlation matrix of NLP features

Compared to the numerical rating features, the NLP-generated features display a much lower inter-correlation. While it may indicate better potential in finding orthogonal signals, it also suggests the potential noisiness of the data generated using NLP techniques.

In general, we found out that comparing the long-term and short-term difference of the *negcount* and *poscount* features as a trading

signal gives a relatively better result, as compared to other approaches such as using long-term averages directly as a signal. While the returns from running the regression analysis framework (same framework as in section 3.2) did not show any statistical significance, the resultant alpha shows strong statistical results with a positive alpha for high bucket (values generated from the signal, in this case, it is the 12-month average of negcount – 6-month average of negcount), and a negative alpha for low bucket, indicating the capturing of abnormal returns based on the CAPM assumptions.

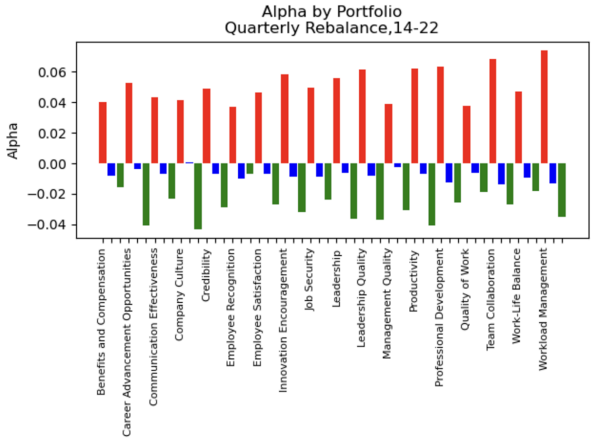

Fig 4.8. Regression alpha on negcount on NLP features

Upon running the regression framework, we also ran the trading signals on our back tester (same as defined in section 3.3). Given the large number of features we are working with, instead of including all the backtests we have conducted through the use of the aggregated NLP dataset, the following section will only contain four of the best alpha signals based on the Sharpe Ratio. Note that while most of the backtests we conducted had a trading period from 2014-01-01 to 2022-12-30, we also include several strategies that have the best performance from 2014-01-01 to 2019-12-30 to exclude the macro impact caused by the COVID-19 pandemic.

1. Benefits and Compensation

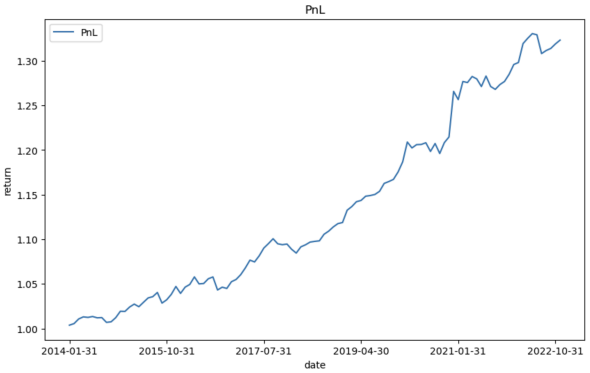

Fig 4.9. B&C Backtest PnL

Signal:
rank(moving_average(B&C_negcount, 6) - moving_average(B&C_negcount, 12))
Trading Period: 2014-01-01 to 2022-12-30
Filters: at least 100 reviews per quarter
Cumulative Return: 33%
Annualized Sharpe: 0.97 (1.09 post COVID)
Average Number of Stocks Traded: 110
Note: Market neutral, decay=6
Description: Assign higher weights to stocks that have had more average number of negative reviews in *Benefits and Compensation* in the past 6 months than in the past 12 months, apply ranking for smoothing.

2. Communication Effectiveness

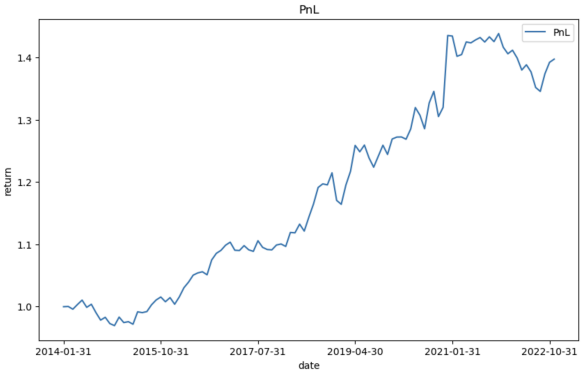

Fig 4.10. CE Backtest PnL

Signal: *rank(moving_average(CE_mean, 6))*
Trading Period: 2014-01-01 to 2022-12-30
Filters: at least 1 review per month
Cumulative Return: 40%

Annualized Sharpe: 0.59 (1.21 post COVID)
Average Number of Stocks Traded: 110
Note: Market neutral, decay=8
Description: Assign higher weight to stocks that exhibit a higher average sentiment score in *Communication Effectiveness* in the past 6 months, apply ranking for smoothing.

3. Professional Development

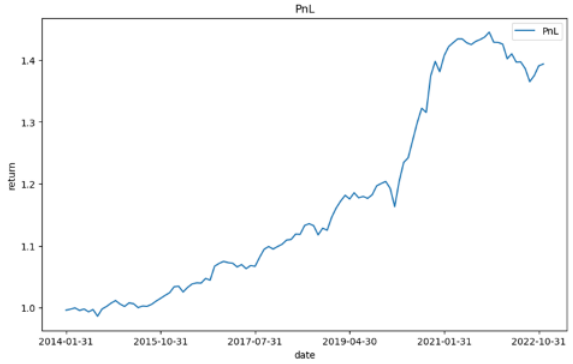

Fig 4.11. PD Backtest PnL

Signal: *rank(moving_average(PD, 1))*
Trading Period: 2014-01-01 to 2022-12-30
Filters: at least 1 review per month
Cumulative Return: 40%
Annualized Sharpe: 0.85 (0.94 post COVID)
Average Number of Stocks Traded: 754
Note: Market neural, decay=6
Description: Assign higher weight to stocks that exhibit a higher average sentiment score in *Professional Development* in the past month, apply ranking for smoothing.

4. Innovation Encouragement

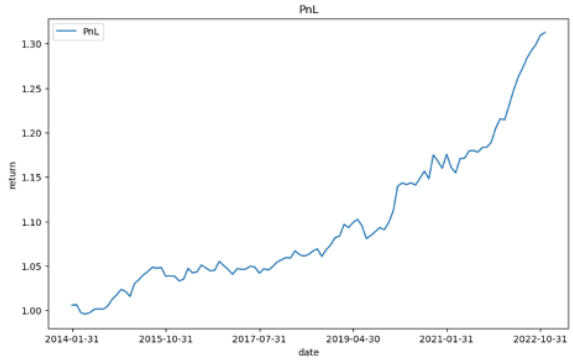

Fig 4.12. IE Backtest PnL

Signal: *moving_average(IE, 6) - moving_average(IE, 3)*
Trading Period: 2014-01-01 to 2022-12-30
Filters: at least 10 reviews per month
Cumulative Return: 31%
Annualized Sharpe: 0.97 (0.35 post COVID)
Average Number of Stocks Traded: 300
Note: Market neutral, decay=0
Description: Assign higher weight to stocks that have had more average number of positive reviews in *Benefits and Compensation* in the past 6 months than in the past 3 months.

Additionally, we plot the correlation heatmap between the monthly returns of the above backtest results. We can see that the correlation between different strategies is very low, suggesting uncorrelated information embedded in these NLP features.

Fig 4.13. Correlation matrix for monthly returns from the previous backtesting results

5. Conclusion

In summary, this project presents a comprehensive analysis of Glassdoor reviews and their relationship with stock performance through exploratory data analysis, regression analysis, and backtesting. Our findings indicate that while overall ratings from

Glassdoor exhibit a Gaussian-like distribution, certain categories, such as Senior Leadership, show distinct patterns that can influence investment strategies.

The regression analysis demonstrates a correlation between higher numerical ratings and abnormal portfolio returns under the CAPM assumptions. The analysis highlights the potential for generating abnormal returns by constructing market-neutral portfolios, in which we conduct backtesting experiments and discover the potential of using employment status as conditional filtering to generate less correlated results in opposition to the direct use of numerical ratings which demonstrated significant correlation and drawdown from 2020-2022.

The integration of textual analysis further enriches our understanding and allows us to identify more orthogonal signals. By consolidating various review categories and employing natural language processing techniques, we uncovered deeper insights into employee sentiments that are less correlated than numerical ratings, suggesting the presence of unique signals that could drive investment decisions. Moreover, the backtesting results reveal that specific trading signals derived from sentiment analysis yield promising cumulative returns and Sharpe ratios. We highlighted review categories such as *Benefits and Compensation, Professional Development, Innovation Encouragement, and Communication Effectiveness*. These insights underline the necessity of incorporating quantitative and qualitative analyses when evaluating company performance.

Overall, this study underscores the value of leveraging employee feedback, both numerical and textual, as a strategic tool for quantitative trading. Future research could expand on these findings by exploring additional sentiment metrics and refining the models to enhance predictive accuracy, group neutralization by industry, and conduct more throughout universe selection techniques, etc, ultimately providing a more robust framework for investment decision-making based on employee sentiment.

References

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Chandola, D., Mehta, A., Singh, S., Tikkiwal, V. A., & Agrawal, H. (2023). Forecasting directional movement of stock prices using deep learning. *Annals of Data Science*, *10*(5), 1361-1378.

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, *6*, 1-25.

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, *34*(7), 3265-3315.

Muthukumar, P., & Zhong, J. (2021). A stochastic time series model for predicting financial trends using nlp. *arXiv preprint arXiv:2102.01290*.

Osterrieder, J. (2023). A primer on natural language processing for finance. *Available at SSRN 4317320*.

Pepe, F., Nardone, V., Mastropaolo, A., Bavota, G., Canfora, G., & Di Penta, M. (2024, April). How do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension* (pp. 370-381).

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49-73.

Appendix

Fig. A.1. QQ plot of numerical rating distributions

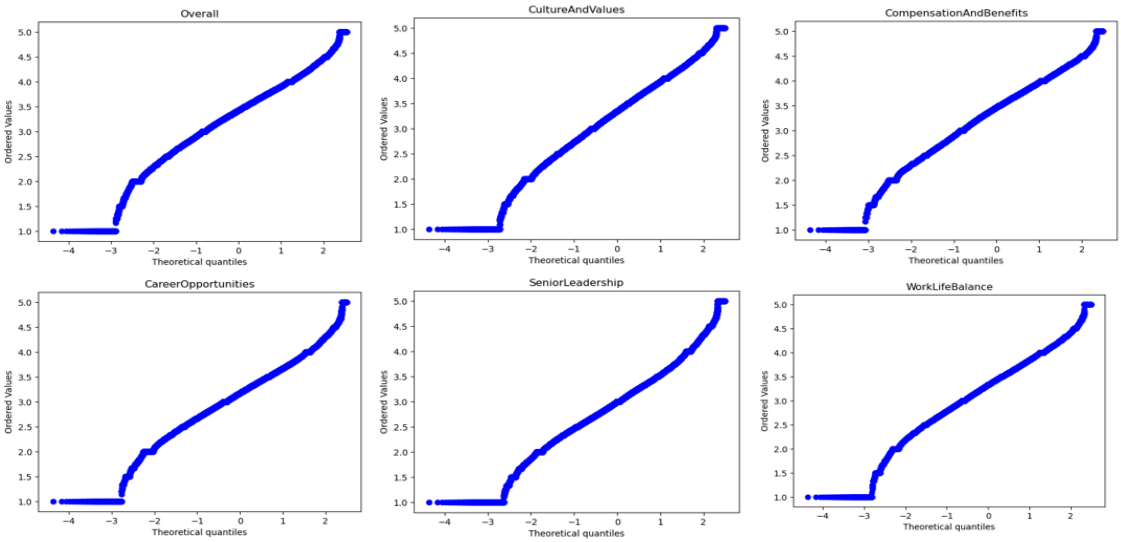

Fig. A.2. QQ plot of NLP features distribution

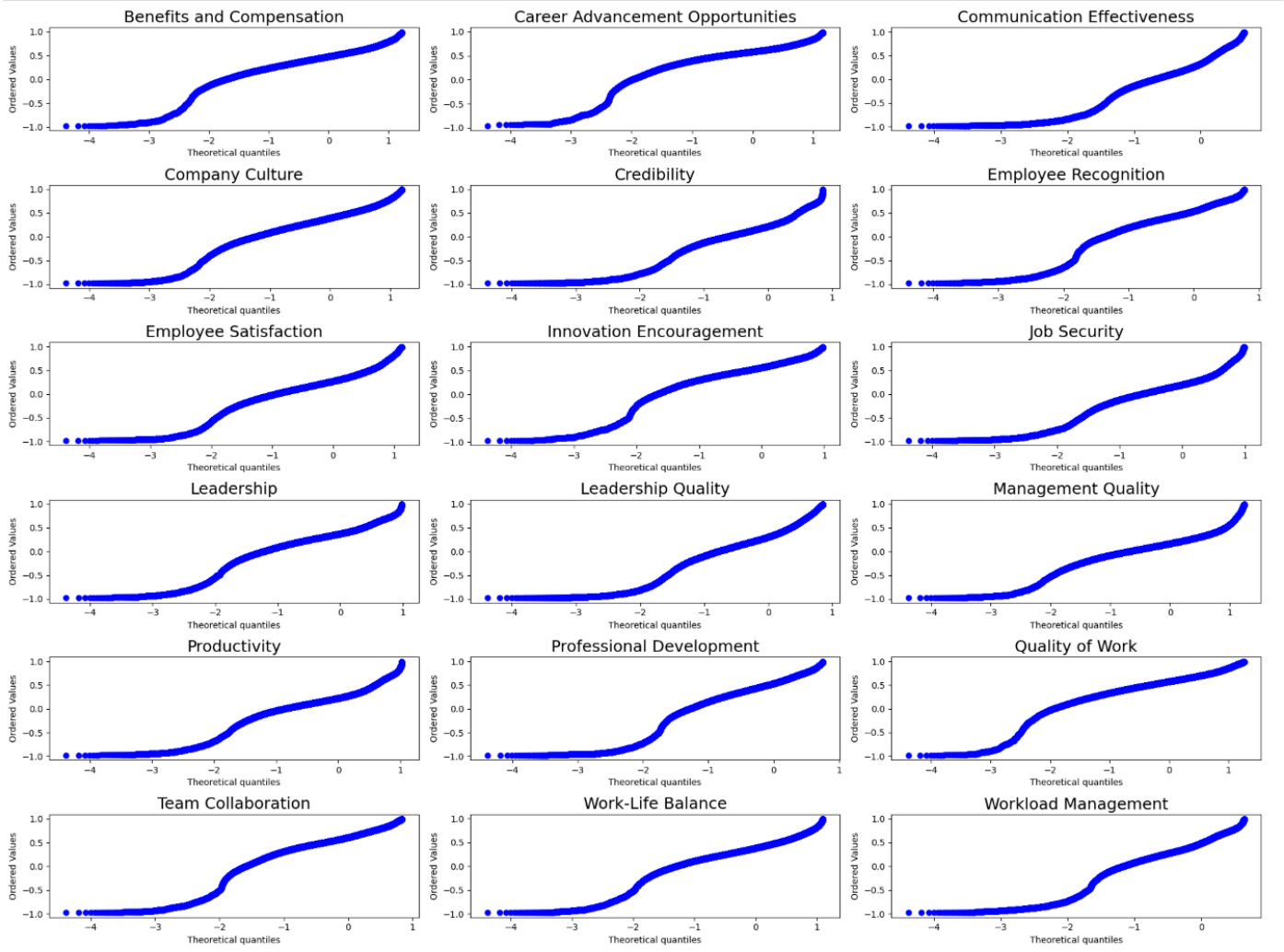

Fig A.3. Distribution of NLP features – only positive sentiments

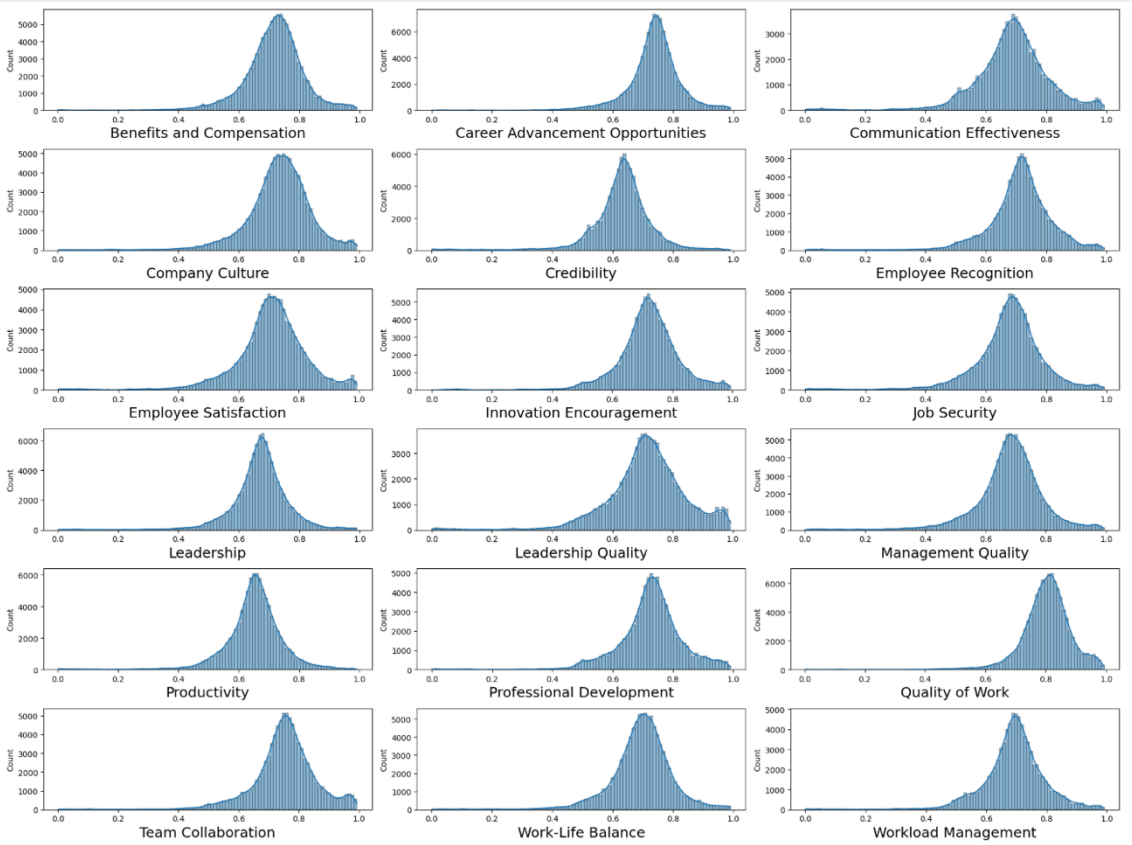

Fig A.4. Distribution of NLP features – only negative sentiment

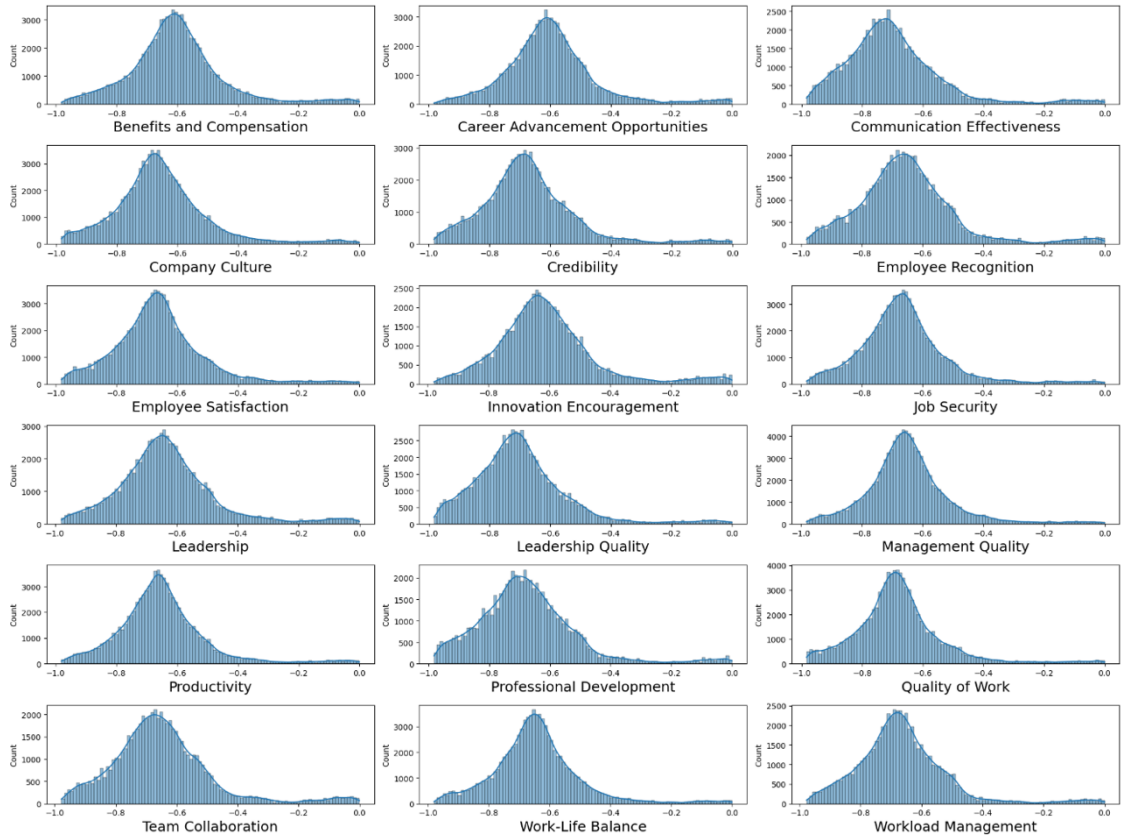

Table A.1. Regression table on numerical ratings

experiment	avg_ret	alpha	t_stat	pval
Overall_high	0.0124	0.0858	32.6696	3.72E-57
Overall_mid	0.0111	0.0347	65.0668	2.81E-87
Overall_low	0.01	-0.0762	46.3563	3.34E-72
CompensationAndBenefits_high	0.0134	0.1255	30.1911	7.14E-54
CompensationAndBenefits_mid	0.0111	0.0213	67.6142	5.28E-89
CompensationAndBenefits_low	0.0088	-0.0478	37.4944	5.39E-63
SeniorLeadership_high	0.0134	0.0936	30.256	5.82E-54
SeniorLeadership_mid	0.011	0.0381	57.9058	4.69E-82
SeniorLeadership_low	0.0093	-0.0632	44.0778	5.38E-70
WorkLifeBalance_high	0.0122	0.0715	32.9774	1.51E-57
WorkLifeBalance_mid	0.0106	0.0376	61.2553	1.44E-84
WorkLifeBalance_low	0.0117	-0.0529	41.4941	2.29E-67
CareerOpportunities_high	0.0131	0.0777	32.6966	3.44E-57
CareerOpportunities_mid	0.0104	0.0499	56.7474	3.74E-81
CareerOpportunities_low	0.0111	-0.096	45.2578	3.76E-71
CultureAndValues_high	0.013	0.0782	32.4863	6.41E-57
CultureAndValues_mid	0.011	0.0395	60.0502	1.11E-83
CultureAndValues_low	0.0096	-0.0575	44.0843	5.30E-70

Table A.2. Regression table on numerical ratings (on unemployed reviews only)

experiment	avg_ret	alpha	t_stat	pval
Overall_high	0.0127	0.1365	19.859	1.65E-37
Overall_mid	0.0105	0.0628	25.4593	5.70E-47
Overall_low	0.0086	-0.0872	22.3966	5.47E-42
CompensationAndBenefits_high	0.0131	0.1957	19.2139	2.57E-36
CompensationAndBenefits_mid	0.01	0.0776	28.3569	2.65E-51
CompensationAndBenefits_low	0.0094	-0.2105	16.0795	3.38E-30
SeniorLeadership_high	0.014	0.1087	19.6238	4.46E-37
SeniorLeadership_mid	0.0104	0.0694	26.597	1.03E-48
SeniorLeadership_low	0.0065	-0.1035	21.3689	3.25E-40
WorkLifeBalance_high	0.011	0.1603	18.0979	3.36E-34
WorkLifeBalance_mid	0.0091	0.048	28.5979	1.20E-51
WorkLifeBalance_low	0.0135	-0.0773	20.6798	5.39E-39
CareerOpportunities_high	0.0108	0.155	20.5398	9.61E-39
CareerOpportunities_mid	0.0108	0.0536	26.139	5.10E-48
CareerOpportunities_low	0.009	-0.0974	19.4408	9.72E-37
CultureAndValues_high	0.0128	0.1485	18.5955	3.75E-35
CultureAndValues_mid	0.0104	0.0536	26.4249	1.88E-48
CultureAndValues_low	0.0077	-0.098	21.8695	4.38E-41

Table A.3. Regression table on NLP features (negative count)

experiment	avg_ret	alpha	t_stat	pval
Benefits and Compensation_negcount_high	0.0121	0.04	41.2393	4.24E-67
Benefits and Compensation_negcount_mid	0.0114	-0.0081	156.5434	3.16E-127
Benefits and Compensation_negcount_low	0.0126	-0.0157	61.6984	6.83E-85
Career Advancement Opportunities_negcount_high	0.0121	0.0525	42.9349	7.53E-69
Career Advancement Opportunities_negcount_mid	0.0116	-0.0039	163.738	2.75E-129
Career Advancement Opportunities_negcount_low	0.0119	-0.0407	70.5537	6.40E-91
Communication Effectiveness_negcount_high	0.0116	0.0435	49.5655	3.78E-75
Communication Effectiveness_negcount_mid	0.0118	-0.0067	153.3633	2.76E-126
Communication Effectiveness_negcount_low	0.0121	-0.0234	81.3324	2.39E-97
Company Culture_negcount_high	0.0121	0.0415	39.3788	4.21E-65
Company Culture_negcount_mid	0.0115	0.0005	171.2908	2.35E-131
Company Culture_negcount_low	0.0123	-0.043	64.3277	9.17E-87
Credibility_negcount_high	0.0119	0.049	44.7319	1.22E-70
Credibility_negcount_mid	0.0119	-0.0067	171.4373	2.15E-131
Credibility_negcount_low	0.0112	-0.0289	70.7919	4.51E-91
Employee Recognition_negcount_high	0.0122	0.0369	41.3335	3.38E-67
Employee Recognition_negcount_mid	0.0118	-0.01	149.0869	5.45E-125
Employee Recognition_negcount_low	0.0114	-0.0067	84.0192	8.04E-99
Employee Satisfaction_negcount_high	0.0117	0.0466	37.8096	2.36E-63
Employee Satisfaction_negcount_mid	0.0118	-0.0067	147.2242	2.05E-124
Employee Satisfaction_negcount_low	0.0118	-0.0266	65.7848	9.04E-88
Innovation Encouragement_negcount_high	0.0125	0.0582	46.3073	3.72E-72
Innovation Encouragement_negcount_mid	0.0114	-0.0087	167.9092	1.93E-130
Innovation Encouragement_negcount_low	0.0124	-0.032	80.4424	7.54E-97
Job Security_negcount_high	0.012	0.0495	39.6195	2.30E-65
Job Security_negcount_mid	0.0118	-0.0086	170.5832	3.64E-131
Job Security_negcount_low	0.0117	-0.0237	71.3371	2.03E-91
Leadership_negcount_high	0.0115	0.0558	46.3922	3.09E-72
Leadership_negcount_mid	0.0117	-0.0065	155.1407	8.17E-127
Leadership_negcount_low	0.0125	-0.0362	75.893	3.25E-94
Leadership Quality_negcount_high	0.011	0.0614	48.9416	1.37E-74
Leadership Quality_negcount_mid	0.0119	-0.008	181.0863	6.61E-134
Leadership Quality_negcount_low	0.0123	-0.0372	74.1657	3.57E-93
Management Quality_negcount_high	0.0124	0.039	38.5998	3.05E-64
Management Quality_negcount_mid	0.0115	-0.0027	144.3622	1.63E-123
Management Quality_negcount_low	0.012	-0.0308	60.9798	2.28E-84
Productivity_negcount_high	0.0118	0.0618	39.0774	9.01E-65
Productivity_negcount_mid	0.0117	-0.007	158.3201	9.60E-128
Productivity_negcount_low	0.0121	-0.0406	66.144	5.15E-88
Professional Development_negcount_high	0.0123	0.063	42.858	9.01E-69
Professional Development_negcount_mid	0.0118	-0.0123	157.7323	1.42E-127
Professional Development_negcount_low	0.0113	-0.0258	89.5207	1.06E-101
Quality of Work_negcount_high	0.0117	0.0378	38.8037	1.81E-64
Quality of Work_negcount_mid	0.0119	-0.0063	150.8571	1.57E-125

Quality of Work_negcount_low	0.0116	-0.0189	62.8656	9.87E-86
Team Collaboration_negcount_high	0.0118	0.068	54.9201	1.07E-79
Team Collaboration_negcount_mid	0.0117	-0.0136	186.4364	3.05E-135
Team Collaboration_negcount_low	0.0121	-0.0271	88.0429	6.06E-101
Work-Life Balance_negcount_high	0.0119	0.047	39.7092	1.83E-65
Work-Life Balance_negcount_mid	0.0119	-0.0095	166.0232	6.37E-130
Work-Life Balance_negcount_low	0.0114	-0.0183	69.5012	3.04E-90
Workload Management_negcount_high	0.0115	0.0739	50.7692	3.29E-76
Workload Management_negcount_mid	0.0119	-0.0129	168.86	1.06E-130
Workload Management_negcount_low	0.0118	-0.0352	79.7311	1.90E-96