Quantitative Trading Strategy using Glassdoor Employee Review

The University of Hong Kong

FINA 4359 Data Analytics, Quantitative Finance, and Blockchain Finance

Group 5 Final Report

Dec 2024

Ku Pak Ho, Alvin (3035949322)

Chu Clement Chiu Wei (3035800796)

See Yat Nam, Harry (3035940168)

---

1. Introduction

In the increasingly competitive financial markets, the use of alternative data has emerged for generating quantitative investment insights. Vast quantities of financial information are produced daily, encompassing formats such as social media posts, regulatory findings, research reports, press releases etc. While some of this data moves markets immediately, other information diffuses more gradually, offering opportunities for deeper insights. By extracting key components and hidden factors from these datasets, it is possible to generate valuable insights and make successful predictions.

This report explores the use of companies' reviews from Glassdoor to construct a market-neutral long-short strategy, aiming to identify company-specific alpha opportunities while neutralizing broader market risks.

The structure of this report begins with the methodology for signal generation (Section 2), followed by the application of the signal to construct a portfolio for alpha generation (Section 3). This is then followed by the presentation of results (Section 4) and concludes with the final summary and insights (Section 5).

2. Signal Generation

Signal generation transforms and processes the raw data from Glassdoor employee review to a convenient format, as well as creates potentially useful features for the later stage of the project. This section includes the processing of numeric ratings (Section 2.1), textual review length (Section 2.2), sentiment analysis (Section 2.3), aspect-based sentiment analysis (Section 2.4), and finally aggregating all the processed data aggregation and grouping them by month (Section 2.5).

2.1 Numeric Ratings

The Glassdoor dataset consists of 7 numeric ratings obtained from employee reviews, namely overall rating, career opportunities, compensation & benefits, culture & values, senior leadership, work life balance, and diversity & inclusion. The distribution is as follows.

	ratingOverall	ratingCareerOpportunities	ratingCompensationAndBenefits	ratingCultureAndValues
count	4.337584e+06	4.337584e+06	4.337584e+06	4.337584e+06
mean	3.524124e+00	2.705236e+00	2.779128e+00	2.679244e+00
std	1.268336e+00	1.773994e+00	1.757067e+00	1.911374e+00
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	3.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
50%	4.000000e+00	3.000000e+00	3.000000e+00	3.000000e+00
75%	5.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00
max	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00

	ratingSeniorLeadership	ratingWorkLifeBalance	ratingDiversityAndInclusion
count	4.337584e+06	4.337584e+06	4.337584e+06
mean	2.473774e+00	2.723920e+00	1.464791e+00
std	1.773788e+00	1.807538e+00	2.035840e+00
min	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.000000e+00	1.000000e+00	0.000000e+00
50%	3.000000e+00	3.000000e+00	0.000000e+00
75%	4.000000e+00	4.000000e+00	4.000000e+00
max	5.000000e+00	5.000000e+00	5.000000e+00

Table 1: Distribution description of 7 numeric ratings

Three additional Glassdoor survey fields of business outlook, ceo approval, and recommend to friend are in the format of ‘positive’, ‘neutral’, ‘negative’, which were mapped to 1, 0, -1 respectively.

	ratingBusinessOutlook	ratingCeo	ratingRecommendToFriend
count	4.337584e+06	4.337584e+06	4.337584e+06
mean	2.071884e-01	2.249072e-01	2.000621e-01
std	6.697530e-01	6.427729e-01	8.220086e-01
min	-1.000000e+00	-1.000000e+00	-1.000000e+00
25%	0.000000e+00	0.000000e+00	-1.000000e+00
50%	0.000000e+00	0.000000e+00	0.000000e+00
75%	1.000000e+00	1.000000e+00	1.000000e+00
max	1.000000e+00	1.000000e+00	1.000000e+00

Table 2: Distribution description of business outlook, ceo approval, and recommend to friend

2.2 Textual Review Length

For the textual review input field from Glassdoor, the length of review could potentially be an indicator of the reviewer's level of engagement and opinion strength, whether positive or negative. Therefore, this feature was extracted for analysis in the later stages.

	summary_length	pros_length	cons_length	advice_length
count	2.820648e+06	2.820648e+06	2.820648e+06	2.820648e+06
mean	2.496283e+01	1.089359e+02	1.712810e+02	6.755599e+01
std	1.877206e+01	1.395294e+02	3.033977e+02	1.474313e+02
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.300000e+01	4.200000e+01	4.300000e+01	0.000000e+00
50%	2.000000e+01	6.500000e+01	7.800000e+01	0.000000e+00
75%	3.200000e+01	1.260000e+02	1.740000e+02	8.500000e+01
max	1.742000e+03	1.793400e+04	2.758900e+04	2.724900e+04

Table 3: Distribution description of review length

Additionally, the number of tokens in each review were also extracted as another feature, which may be useful for reducing noise in data. The methodology of counting tokens consists of NLP techniques in tokenization and removal of stop words and punctuations.

	summary_token_count	pros_token_count	cons_token_count	advice_token_count
count	2.820648e+06	2.820648e+06	2.820648e+06	2.820648e+06
mean	2.973558e+00	1.085243e+01	1.589832e+01	6.244680e+00
std	1.847269e+00	1.259211e+01	2.694033e+01	1.322215e+01
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.000000e+00	5.000000e+00	5.000000e+00	0.000000e+00
50%	3.000000e+00	7.000000e+00	8.000000e+00	0.000000e+00
75%	4.000000e+00	1.200000e+01	1.600000e+01	8.000000e+00
max	1.490000e+02	1.694000e+03	2.494000e+03	2.300000e+03

Table 4: Distribution description of review token length

2.3 Sentiment Analysis

The Glassdoors textual reviews came uncleaned with HTML tags, machine generated symbols and markdown language, which are unwanted noise and cleaned away using regular expressions. After that, the cleaned text in each column in ‘summary’, ‘pros’, ‘cons’, ‘advice’ was separately analyzed to determine the sentiment, classifying text into positive, neutral or negative sentiment with a confidence score. In this step, the model chosen for the sentiment analysing task was RoBERTa. It leverages the transformer-based deep learning model, which is more capable in understanding complex context compared to traditional rule based models.

Example user review	RoBERTa output	Vader output
Look like a Gap ad or want "just a job"? Apply here. Want fulfilment in your career? Look elsewhere.	Neutral	Positive
Although it's a great place to work for a big company, it is a big company. Too many people making too few decisions. Far too many people I know spend their time working and reworking and reworking powerpoint slides that can be reviewed, chewed on and re-edited again to show to someone else. This lack of single point of accountability was really bad. Also, with so many people career advancement is getting to be pretty difficult. It's getting very tenure based, like a Proctor and Gamble. This is a bummer. Finally, I think that too few people really have the technical and customer passion that's needed there. Few people are really up to speed on the latest products coming out from competitors or how our customers interact with them. This is a bad habit and too many people have gotten lazy.	Negative	Positive

Table 5: Example sentiment analysis using RoBERTa and Vader

reviewId	summary_sentiment	pros_sentiment	cons_sentiment	advice_sentiment
49	0.9581	0.9514	0.0000	0.988652
83	0.0000	0.949141	-0.825076	-0.755065

Table 6: Example dataframe output generated from RoBERTa sentiment analysis

	reviewId	summary_sentiment	pros_sentiment	cons_sentiment	advice_sentiment
count	4.320624e+06	4.320624e+06	4.320624e+06	4.320624e+06	4.320624e+06
mean	3.927943e+07	2.625404e-01	6.988271e-01	-3.890370e-01	1.781537e-02
std	2.659047e+07	5.771621e-01	4.159195e-01	4.851078e-01	3.365856e-01
min	4.900000e+01	-9.848647e-01	-9.849299e-01	-9.854066e-01	-9.846737e-01
25%	1.361665e+07	0.000000e+00	6.544351e-01	-8.115045e-01	0.000000e+00
50%	3.824138e+07	0.000000e+00	9.075719e-01	-5.673188e-01	0.000000e+00
75%	6.403563e+07	8.891091e-01	9.616914e-01	0.000000e+00	0.000000e+00
max	8.542543e+07	9.939519e-01	9.940696e-01	9.934777e-01	9.939177e-01

Table 7: Distribution description of the RoBERTa sentiment analysis

The tables above provide examples of input and output data. For simplicity, the output logits from the sentiment analysis model are mapped to a range between -1 and 1. Higher scores indicate a greater probability of positive sentiment and vice versa, while a score of 0 represents neutral sentiment.

2.4 Aspect-based Sentiment Analysis

Apart from sentiment analysis, there are more features that can be extracted from the text data, for example topic modelling techniques such as LDA and BERTopic can be used to identify hidden patterns of topics by grouping related words into themes. Despite those being common methods, this project tested another approach with large language models. First, GPT-4o was used to generate a list of related words for each topic. This generated dictionary was then reviewed and validated manually. For those employee reviews with any word match with the dictiontion, the sentiment towards the topic category was analyzed using Llama-3.2-3B-Instruct.

Innovation	Integrity	Quality	Respect	Teamwork
innovative	integrity	quality	respectful	teamwork
cutting-edge	honesty	excellent	polite	team
inventive	upright	superior	considerate	collaboration
creative	truth	caliber	open	engagement
ideation	wholeness	standard	trust	cooperation
imaginative	compliance	best	esteemed	coordination
novel	legal	expectations	regarded	communication
visionary	accountability	high-end	revered	helpful
entrepreneurial	responsibility	refined	honorable	synergy
technology	diligence	durable	admirable	aligned
digital	reliability	decent	appreciative	interaction
modernized	conduct	superb	deferential	unity
digitalization	law	top	venerable	unified
modern	governance	meritorious	dignity	collective
futuristic	fiduciary	distinctive	courtesy	solidarity
advanced	authenticity	worthy	regardful	people
futurism	sincerity	prestigious	civility	camaraderie
pioneering	moral	classy	gracious	supportive
disruption	ethical	eminence	admire	coworkers
transformative	ethos	eminent	disabled	contribution
disruptive	incorruptibility	perfect	tolerance	shared
revolutionary	rectitude	reliable	conflict	responsibility
breakthrough	probity	premium	inclusion	partnership
groundbreaking	principle	benchmark	recognition	group
trailblazing	righteous	high-quality	diversity	assist
state-of-the-art	decency	first-class	equity	mutual
forward-thinking	virtue	top-notch	dei	jointly
change	assurance	superlative	empathy	joint
future	candor	outstanding	respectable	collaborator
agile	candid	world-class	diverse	interdependence
startup	fair	best-in-class	professional	cohesive
create	justice	flawless	discrimination	peer
prototype	trustworthy	unmatched	lesbian	leadership
design	transparent	optimal	gay	cohesion
iteration	veracity	finest	bisexual	discuss
renovation	conduct	satisfaction	queer	interact
evolution	diligent	satisfy	friendly	
new	control	exceptional	thoughtful	
ingenuity	correct	product		
	lawful	service		
	right	kpi		
	licit	meticulous		
	legitimate			
	permission			
	valid			
	legit			
	illegal			
	criminal			
	terrorist			
	unlawful			

Table 8: Dictionary for each topic

	Innovation	Integrity	Quality	Respect	Teamwork
% of reviews matched	13.8%	9.4%	28.3%	11.3%	42.4%
Number of reviews matched	594774	407445	1220682	487638	1832236
Total number of reviews	4320624	4320624	4320624	4320624	4320624

Table 9: Match rate between dictionary and reviews

This dictionary was designed for interpretability. The word listed was first created using GPT-4o and the manual validation step refined the dictionary by eliminating ambiguous terms, ensuring the words within the list are likely to be relevant to the topic. Then the dictionary was used to match each review after NLP preprocessing techniques, including tokenization, stemming and lemmatization with part of speech tagging.

Besides, the topic categories (i.e. innovation, integrity, quality, respect, teamwork) were inspired by Li et al (2020), which were proven to be correlated with business outcomes. Also, this combination of topics complements the existing data because some other aspects were already obtained directly from the reviewer as a Glassdoor numeric input field, which includes ratings of career opportunities, compensentions, etc. Therefore, topics highly relevant to the existing features were not created to avoid redundant features and overfitting.

The matched reviews were then fed into the Llama-3.2-3B-Instruct to evaluate sentiment towards each specific topic. The prompt used was as follows.

You are an Aspect-Based Sentiment Analysis assistant.
Your task is to analyze the sentiment of text concerning a specific aspect: '{aspect}'.

For each review provided by the user, you will give a sentiment score between 0 and 1 for the aspect '{aspect}'.
The score should reflect how positive or negative the review is specifically towards '{aspect}', where:
- 0 means very negative
- 0.5 means neutral
- 1 means very positive

If the '{aspect}' is not mentioned in the review, return a score of 0.5.
Please respond only with a number between 0 and 1, rounded to two decimal places

	reviewId	innovative_llama	integrity_llama	quality_llama	respect_llama	teamwork_llama
count	2.820648e+06	594774.000000	407445.000000	1.220682e+06	487638.000000	1.832236e+06
mean	3.699974e+07	0.492772	0.459785	5.303345e-01	0.485621	5.209131e-01
std	2.691232e+07	0.150095	0.228976	2.319804e-01	0.209551	2.012196e-01
min	4.900000e+01	0.000000	0.000000	0.000000e+00	0.000000	0.000000e+00
25%	1.152852e+07	0.500000	0.430000	5.000000e-01	0.500000	5.000000e-01
50%	3.322254e+07	0.500000	0.500000	5.000000e-01	0.500000	5.000000e-01
75%	6.223808e+07	0.500000	0.500000	6.700000e-01	0.500000	6.700000e-01
max	8.542543e+07	0.990000	0.990000	9.900000e-01	0.990000	9.900000e-01

Table 10: Distribution description of the Llama aspect based sentiment analysis

2.5 Data Aggregation and Grouping

All the aforementioned features were aggregated and grouped by month by calculating the mean value corresponding to the same company within the same month. Similarly, for the U.S. equities return data, the original daily returns from the dataset were used to compute the monthly returns, which were then aggregated on a one-month forward basis.

2.6 Features Scaling

A scaler with the following formula was applied to the dataset, after clipping the data to a specified range.

Clip the data to the range between the 5th and 95th percentiles.
X_scaled = (X - median) / IQR where IQR = Q3 - Q1

This transformation centers the data around 0 with the same scale to ensure equal distribution of each feature, while being robust to outliers.

The boxplot of the data before and after applying the scaler is shown below.

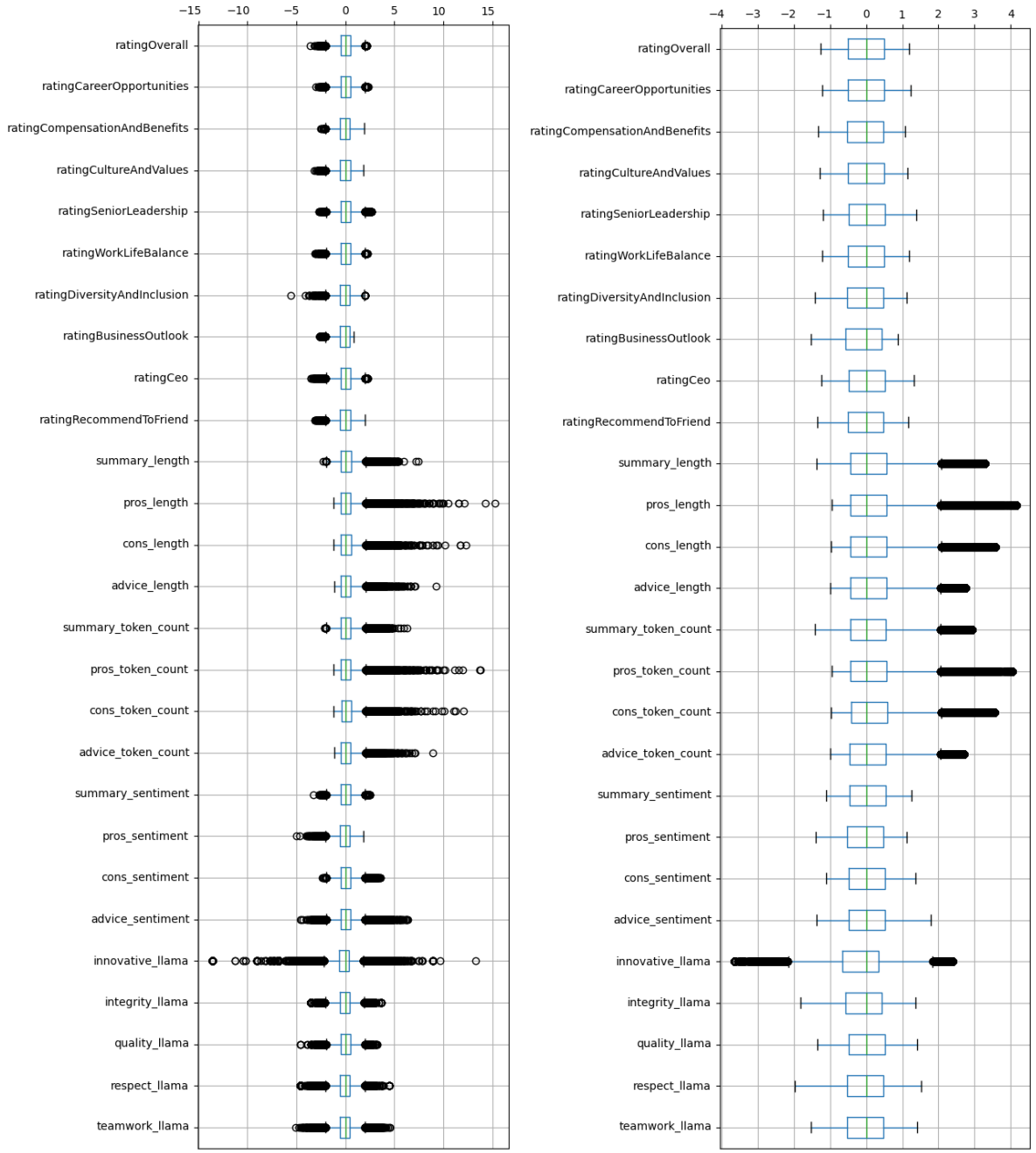


Figure 1: Boxplot before scaling (left), Boxplot after scaling (right)

Due to the high number of outliers in innovative_llama, this feature was excluded to minimize noise and improve data quality.

2.7 Features Selection

Feature selection aims to remove redundant features that negatively impact the model performance. To find out features with high correlation, a heatmap of features correlation was used.

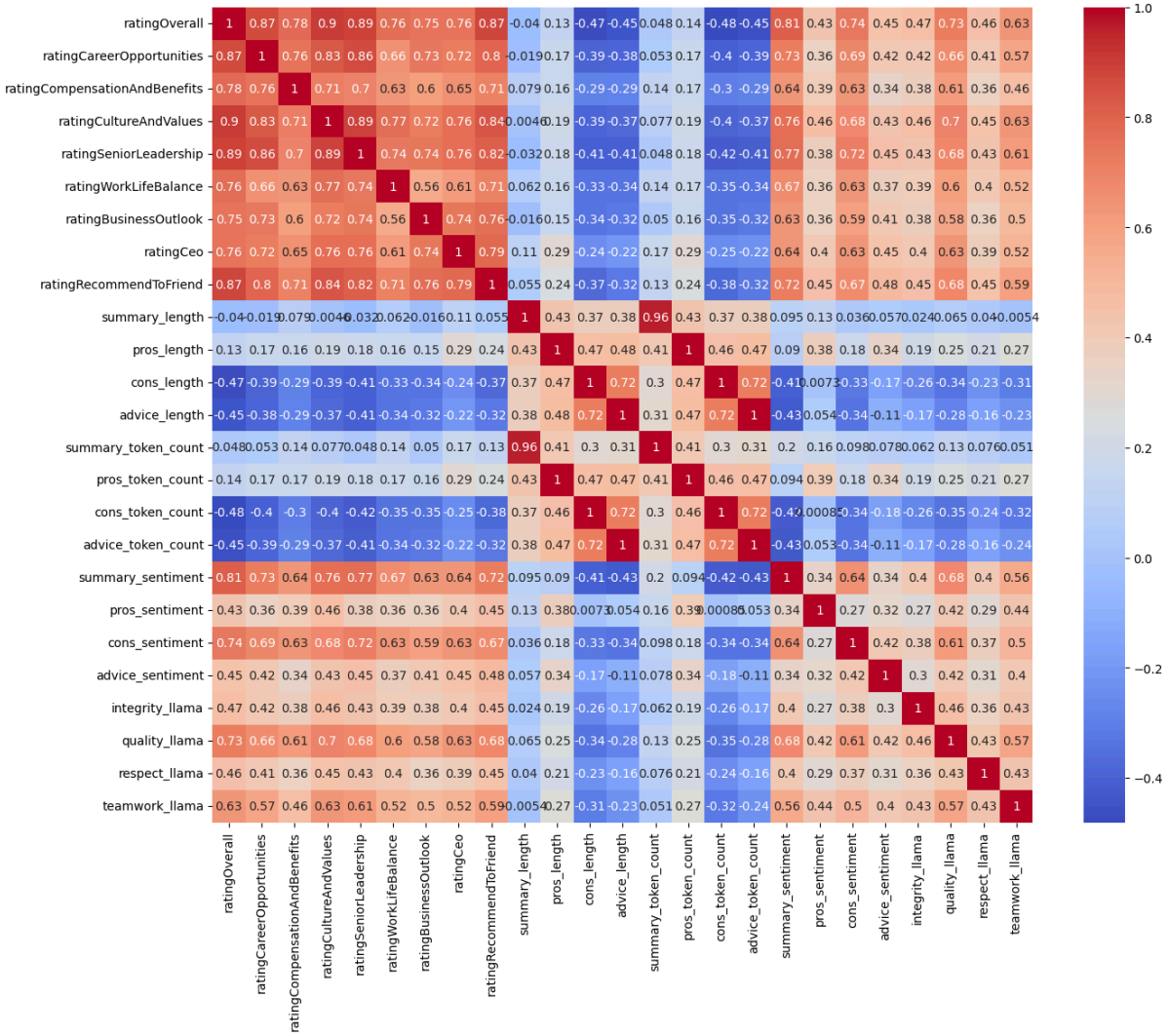


Figure 2: Heatmap of features correlation

Given the information from correlation heatmap, the following steps is taken:

- Since the review length and review token count has correlation close to 1, columns with review length were removed.
- A relatively high correlation was observed within all ratings features (0.7-0.9), these columns were combined using simple mean.

Then, these features are left as shown in the correlation heatmap.

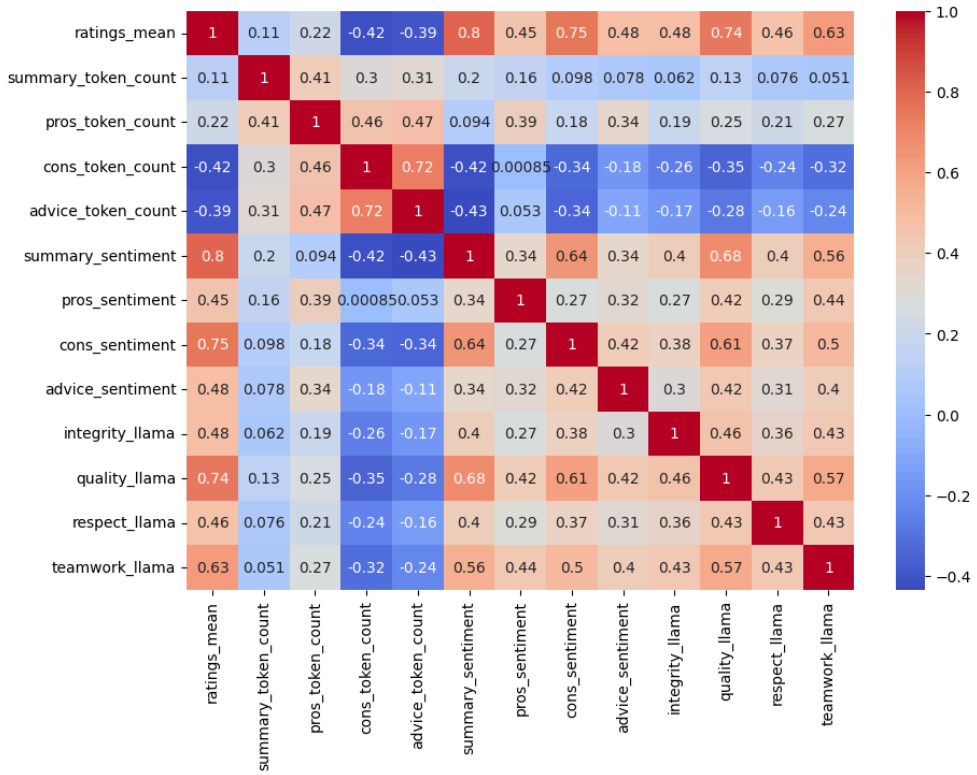


Figure 3: Heatmap of features correlation (after processing)

Finally, the mean of the above features was calculated to generate the trading signals, resulting in the following distribution. The spike in frequency at the rightmost end of the distribution curve could be attributed to the presence of outliers in features, as illustrated in Figure 1.

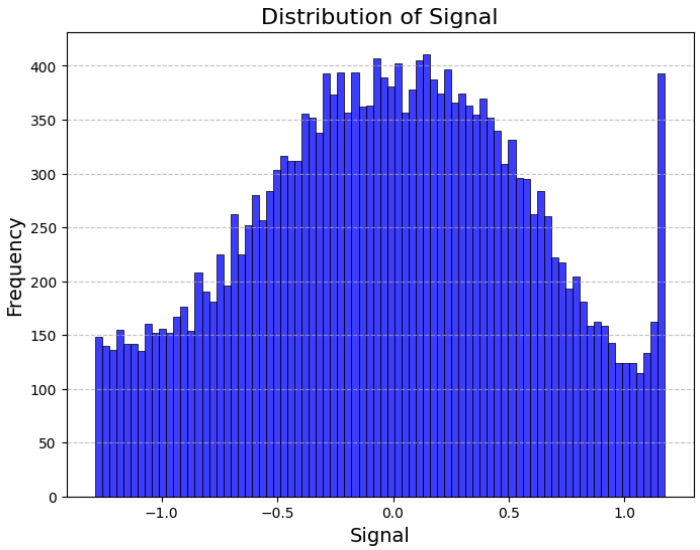


Figure 4: Distribution of signal frequency against signal strength

3. Alpha Generation

The system generates alphas from the signals to create market neutral portfolios to be backtested. The strategies rebalance weights every month, using the monthly aggregated signal mean from Section 2. The neutralization procedures can be further customised through different operators to reduce turnover, industry-wide co-movements or other factors unrelated to the alpha signal. In the following sections, details of the signal-to-portfolio operators are discussed.

3.1 Portfolio Neutralization Methods

Portfolio Neutralization methods provide a simple, non-parametric approach to converting raw signals into positions. While more sophisticated approaches including machine learning can learn to optimize portfolios given a set of signals, they are prone to overfitting in such a small dataset.

We chose to not use the approach of portfolio sorting, i.e. the approach of bucketing stocks by the signal value, and long the top bucket while shorting the bottom bucket. This is since the approach would drop all the stocks in the middle buckets, thus shrinking the tradeable universe even smaller.

The signal extraction methods in Section 2 yields an aggregated score for each stock at each month. The signal output is represented in the form of a 2 dimensional matrix $S \in R^{t\,X\,s}$, where t is the number of months, and s is the number of stocks in the universe. The objective of portfolio neutralization methods is to convert the signal matrix $\boldsymbol{S}$ into weight matrix $W \in R^{t\,X\,s}$, ***where each row sums up to 0 and each row's absolute sum equals 1***. This ensures the portfolio weights of each month represents a market neutral portfolio with consistent scaling.

3.1.1 Overall Market Neutralization

Overall Market Neutralization performs neutralization on the entire universe in each month. Essentially, it assigns a more positive weight to stocks with a greater signal value, and a more negative weight to stocks with a smaller signal value. The strategy would long stocks with a greater signal value, while short stocks with a smaller value. The method is defined as follows:

Considering matrix S:
For each month t:
Minus each stock's signals by the mean of signals of <u>all stock</u> in month t

Divide the results by the resulting absolute sum of signals of all stocks in month t

Output matrix W

The resulting matrix satisfies the condition that each row sum up to 0 and each row's absolute sum equal 1.

This method is simple and easy to implement. However, it neutralizes signals by the entire market, which may fail to take into account the effects of sector/industry trends. To further isolate the alpha from the Glassdoor data, we attempted another approach utilizing group classification in neutralization.

3.1.2 Market Neutralization by group classification

Market Neutralization by group classification further isolates the alpha signals from sector or group-specific risk. We test neutralization on three group classifications: sector, industry, and subindustry grouping. The method is defined as follows:

Considering matrix S:

For each month t:

Minus each stock's signals by the mean of signals of all stock in the same group in month t

Divide the results by the resulting absolute sum of signals of all stocks in the same group in month t

Output matrix W

3.2 Weight Smoothing

Weight smoothing refers to applying rolling averages or other smoothing techniques to weights of stocks over time. We apply weight smoothing for two reasons: reducing the turnover of the strategy, and incorporating past signal information in today's weights.

High turnover is undesirable for a strategy in real trading scenarios, as it increases transaction costs of rebalancing the portfolio. It can also enable the weights to incorporate time-dependent information. As the signal for each month is calculated by reviews in that month, it may not be able to reflect the sentiment of reviews from prior months. For instance, if a company receives 1 positive review in the

current month, but it has 3-4 negative reviews in the past few months, then using this month's review alone cannot fully reflect the sentiment of the stock. The method is as follows:

1. *Conduct neutralization on signals (overall or group)*
2. *Smooth each stock's weights in an exponential/simple moving average of N months*
3. *Conduct neutralization on the results*

3.3 Backtest Design

Given the weight matrix ***W*** and a returns matrix ***R*** of the same shape representing the return of each stock at each month, we calculate the strategy returns and sharpe ratio as follows:

1. *Perform an element-wise multiplication of **W** and **R** to obtain a new matrix of the same shape.*
2. *Sum up the values in each row in obtain a returns vector $V \in R^t$, representing the strategy return in each month*
3. *Calculate Sharpe ratio as the mean of values in **V** divided by the standard deviation of values in **V***

We also compare the number of stocks in the portfolio with non-zero weights and maximum absolute weight across time. Ideally, the more diversified the portfolio, the less influence one stock can have on the strategy performance, hence less idiosyncratic risk.

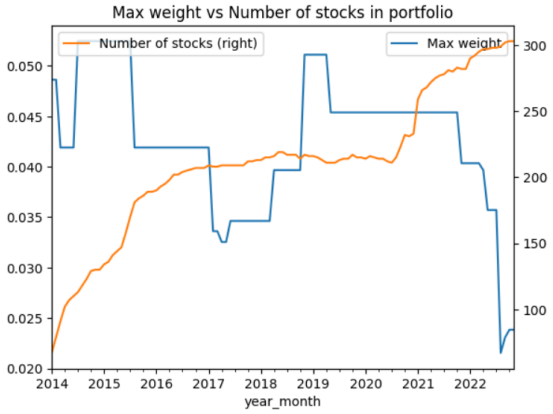
4. Results & Findings

4.1 Detailed Results

We categorized companies' market capitalization into three distinct groups: small-cap (less than $2 billion), mid-cap ($2–10 billion), and large-cap (greater than $10 billion). The subsequent graphs present the results of 20 distinct strategies, stratified by market capitalization and adjusted using the aforementioned neutralization methods, including sector, industry group, industry, and sub-industry levels.

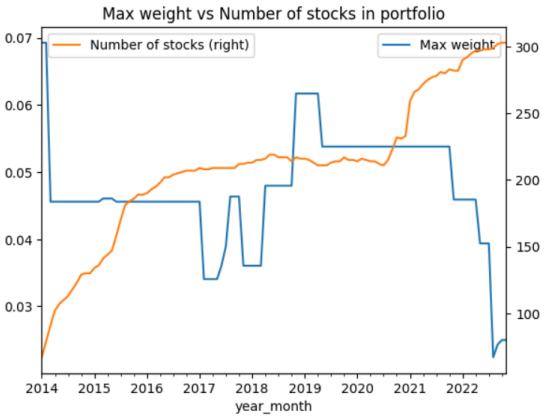
Large cap (over 10 billion), neutralize by sector

Overall Annual Sharpe Ratio: 0.8119154145296228



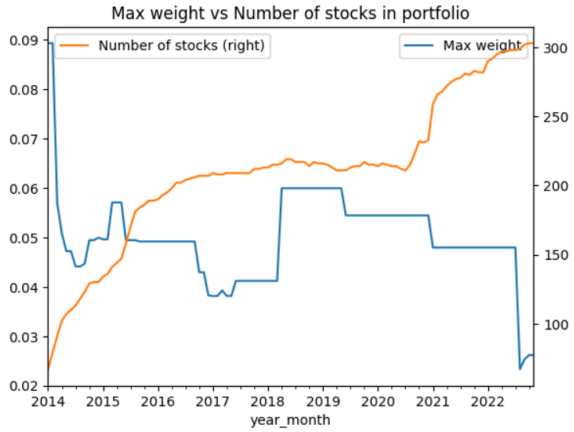
Large cap (over 10 billion), neutralize by industry group

Overall Annual Sharpe Ratio: 0.8381637673051698



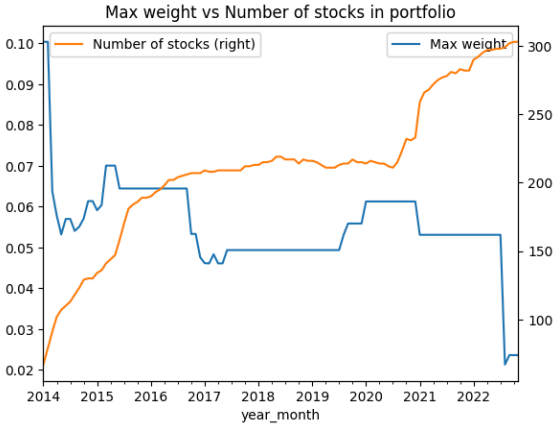
Large cap (over 10 billion), neutralize by industry

Overall Annual Sharpe Ratio: 0.8478585565167795



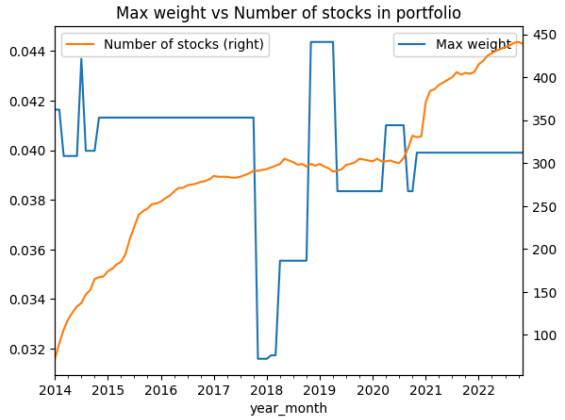
Large cap (over 10 billion), neutralize by sub-industry

Overall Annual Sharpe Ratio: 0.8811544729409291



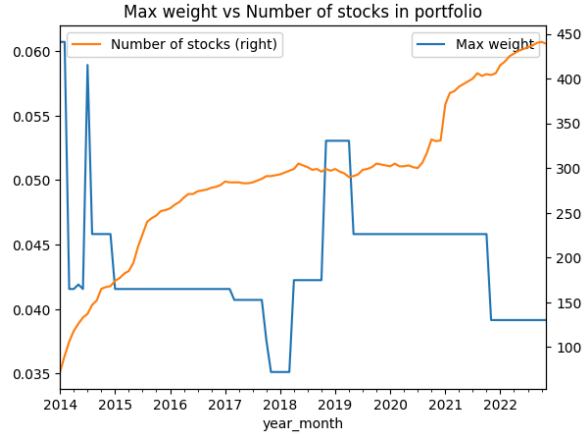
Mid cap (2 - 10 billion), neutralize by sector

Overall Annual Sharpe Ratio: 0.7140988934696908



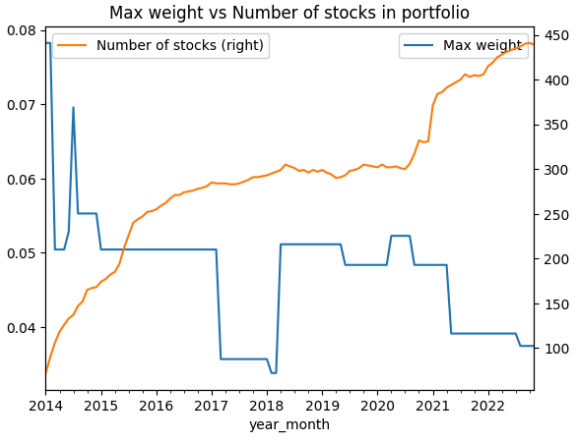
Mid cap (2 - 10 billion), neutralize by industry group

Overall Annual Sharpe Ratio: 0.677656999042748



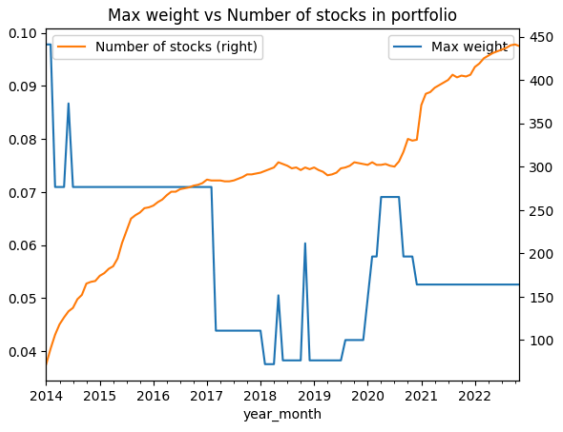
Mid cap (2 - 10 billion), neutralize by industry

Overall Annual Sharpe Ratio: 0.6852475573957533



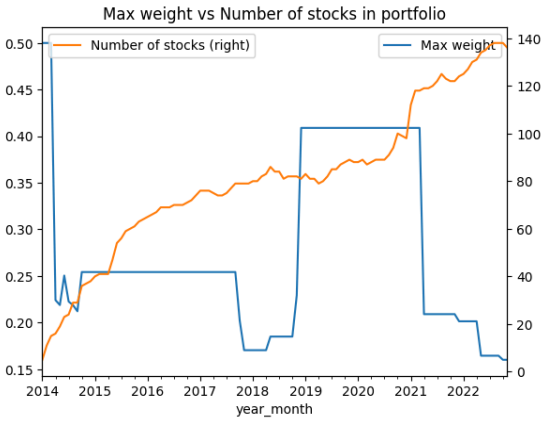
Mid cap (2 - 10 billion), neutralize by sub-industry

Overall Annual Sharpe Ratio: 0.6508670904609668



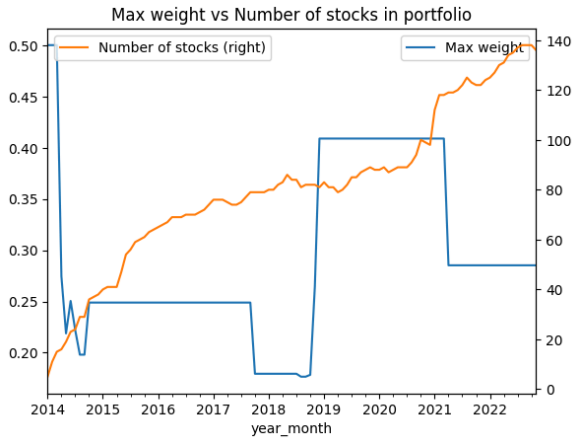
Small cap (under 2 billion), neutralize by sector

Overall Annual Sharpe Ratio: 0.26054039322278605



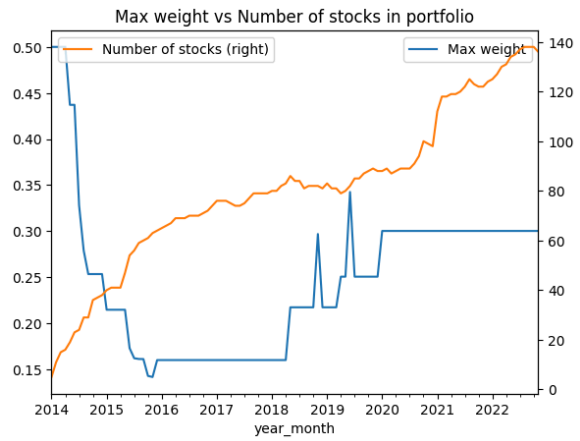
Small cap (under 2 billion), neutralize by industry group

Overall Annual Sharpe Ratio: 0.2730302089064921



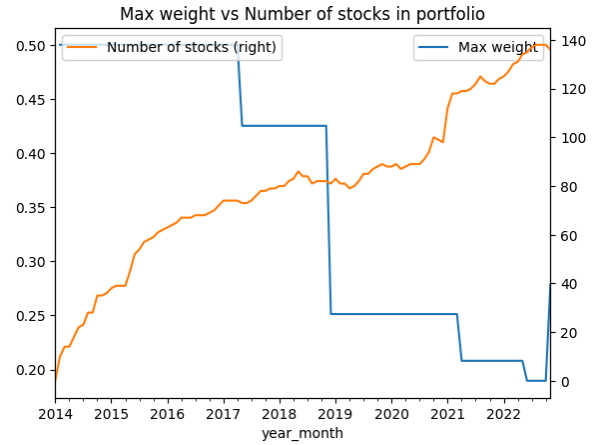
Small cap (under 2 billion), neutralize by industry

Overall Annual Sharpe Ratio: 0.39245144811363664



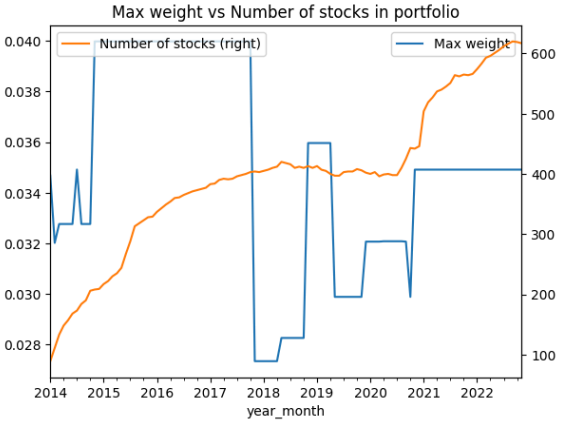
Small cap (under 2 billion), neutralize by sub-industry

Overall Annual Sharpe Ratio: 0.6715483395903372



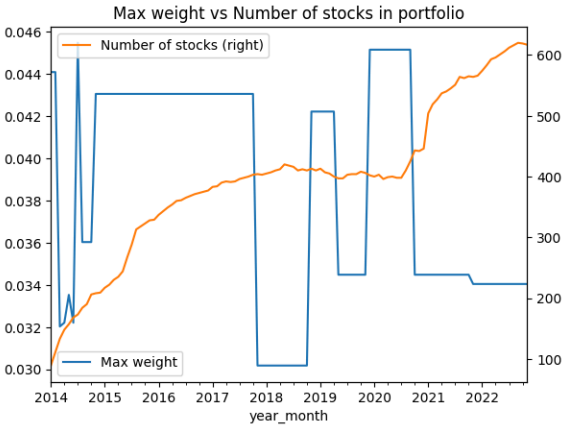
Any market cap, neutralize by sector

Overall Annual Sharpe Ratio: 0.6885293389149663



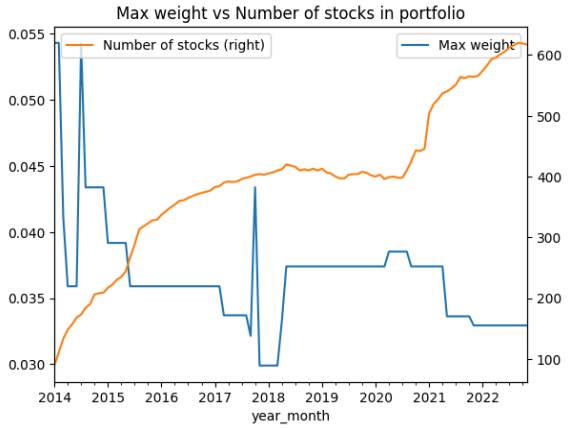
Any market cap, neutralize by industry group

Overall Annual Sharpe Ratio: 0.7046021529428317



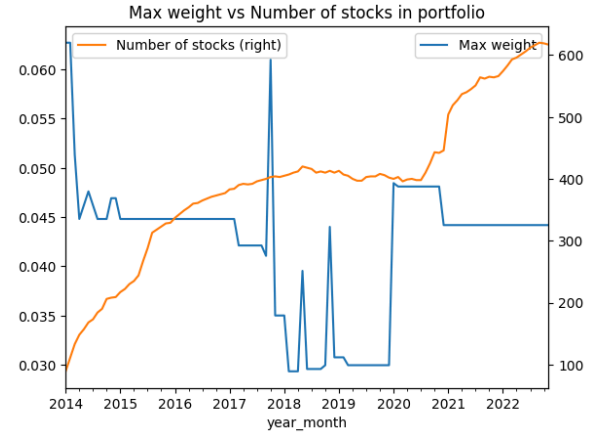
Any market cap, neutralize by industry

Overall Annual Sharpe Ratio: 0.69648500970635



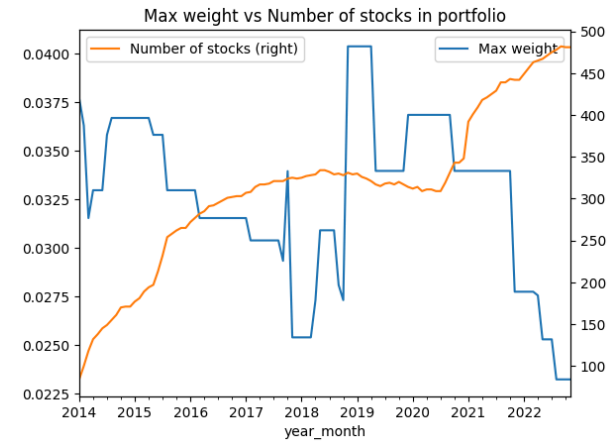
Any market cap, neutralize by sub-industry

Overall Annual Sharpe Ratio: 0.773730548969548



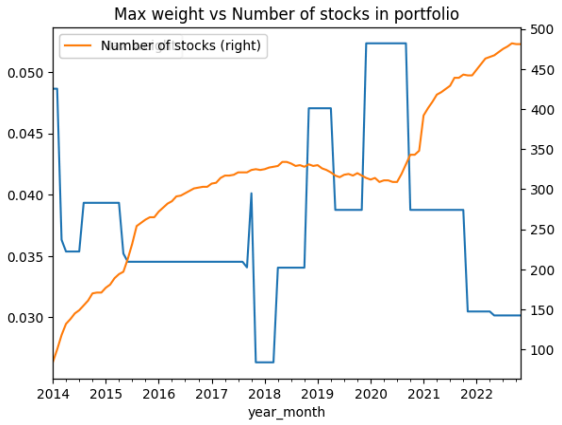
Mid and large cap (over 2 billion), neutralize by sector

Overall Annual Sharpe Ratio: 0.760714700660661



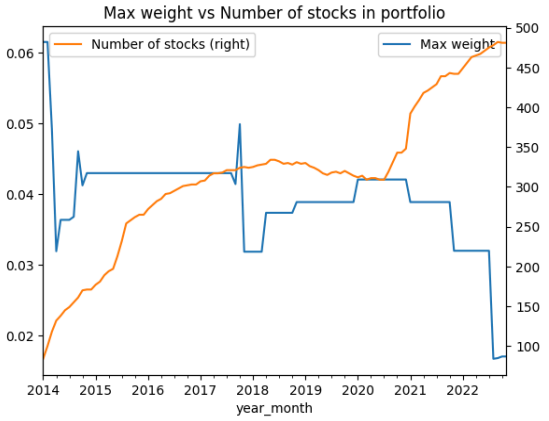
Mid and large cap (over 2 billion), neutralize by industry group

Overall Annual Sharpe Ratio: 0.8274409487280746



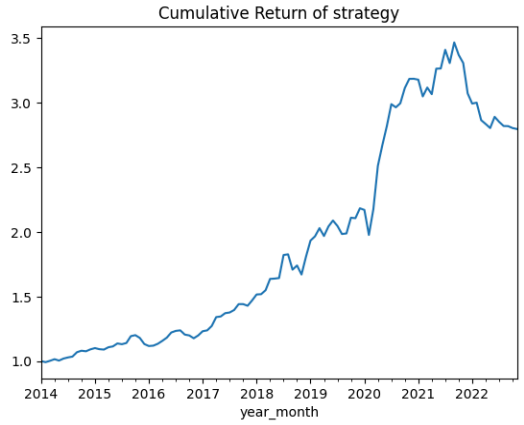
Mid and large cap (over 2 billion), neutralize by industry

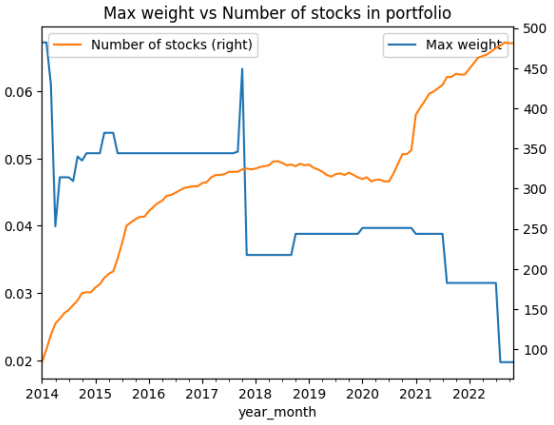
Overall Annual Sharpe Ratio: 0.7754633421759592



Mid and large cap (over 2 billion), neutralize by sub-industry

Overall Annual Sharpe Ratio: 1.0171623591961751





4.2 Results Summary

Sharpe ratio	Neutralize by sector	Neutralize by industry group	Neutralize by industry	Neutralize by sub-industry
Large cap (over $10 billion)	0.8119	0.8382	0.8479	0.8812
Mid cap ($2 - $10 billion)	0.7141	0.6777	0.6852	0.6508
Small cap (under $2 billion)	0.2605	0.2730	0.3925	0.6715
Any market cap	0.6885	0.7046	0.6965	0.7737
Mid and large cap (over $2 billion)	0.7607	0.8274	0.7755	1.017

Table 11: Strategies Sharpe Ratio Summary

5. Conclusion

This paper investigated the potential of Glassdoor employee reviews as an alternative data source for generating quantitative trading strategies. By transforming both numeric ratings and textual reviews into actionable signals through large language models, we constructed market-neutral long-short portfolios aimed at capturing company-specific alpha.

Backtesting results show a modest performance with annualized Sharpe ratios ranging from 0.7 to 1 across different market capitalizations and different neutralization methods. Notably, this strategy works the best on mid and large-cap stocks while performing market neutralization within sub-industry, exhibiting the highest Sharpe ratio of 1.017. This finding suggests that employee reviews from Glassdoor contain predictive information for stock returns to a small to moderate extent.