Phase III: The effect of public sentiment on company value

Ali Khambati, Jackson Joffe, Nimedia Ozinegbe, and Donn Boddie

December 5, 2021

1 Research Question

Does public opinion of a company correlate with its stock returns?

2 Background

Recently, especially on the heels of a pandemic, the use of social media has increased drastically. Pew Research estimates a little over three out of every ten Americans are "almost constantly" online.[1] While Instagram and Facebook remain popular among their core users, Twitter has seen explosive growth as a means for users to communicate information about companies, current events and the government. It is estimated that almost half of Americans aged 18-29 use Twitter as of February 2021.[2] Even amongst more seasoned traders,

[1] https://www.pewresearch.org/fact-tank/2021/03/26/about-three-in-ten-u-s-adults-say-they-are-almost-constantly-online/

[2] https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/

like hedge fund managers, scraping public social media sites has become a core component of their valuation methodology. Surveys by AIMA and SS&C have shown that over 69% of hedge fund managers utilize alternative data, which includes social media scraping, to augment traditional long-short strategies.[3] Examining market sentiment is rooted in the principles of corporate finance. According to the 'expectations treadmill' principle of asset performance, the return an investor can earn on an asset is not the same as the company's own return on invested capital.[4] Fundamentally, individual groups of investors will earn different returns because they pay different prices for the shares based on their expectations of future performance. Public sentiment around a company or asset can serve as a proxy for these future expectations, and this in turn can help investors allocate capital efficiently.

With this background in mind, we aim to analyze the effect of public sentiment, using Tweets as a proxy, on the daily returns of top companies. We want to analyze the correlation of this sentiment as well as other factors driving the daily returns of different investments.

We believe there are a couple of factors that warrant exploration in our analysis.

1. **Previous Research:** a study by Mao et al. found that financial decision-making is dependent on mood and emotion[5]. Casual links have been established between finance and sentiment; therefore we want to focus this paper specifically on stock pricing.

2. **Herd Mentality:** As Moldovan's 2010 article notes, investors tend to follow popular advice.[6] In an age of virality, we want to analyze whether public sentiment can snow-

[3]https://www.hedgeweek.com/2020/05/04/285283/hedge-funds-use-alternative-data-tipped-surge-new-industry-study-finds

[4]https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-ceos-guide-to-corporate-finance

[5]https://arxiv.org/abs/1010.3003

[6]https://ideas.repec.org/a/aio/rteyej/v1y2010i15sp21-26.html

ball and drive daily return fluctuations or whether positive sentiment is a product of multiple independent events.

3. **Efficient Market Hypothesis:** A frequently studied hypothesis in financial economics, the efficient-market hypothesis states that asset prices reflect all available information.[7] With this in mind, we want to analyze whether asset prices are truly reflective of public sentiment, another form of available information.

3 Initial Hypotheses

Conventional research shows that market sentiment is somewhat related with market volatility. Our initial hypothesis extends this idea, relating market sentiment to daily return movements; we hypothesize that sentiment is directly correlated with market value. This hypothesis is motivated by behavioral finance literature, which shows that financial decision-making is dependent on emotions and investors are quick to follow popular advice. Anecdotally, meme stocks like Gamestop enjoyed historical price appreciation largely because of collective movements that were fostered through social media. The alternative hypothesis is that market sentiment is not correlated with market value. We also would like to explore other factors that effect this relationship including:

1. **Volume Composition:** We believe if a stock's volume consists more of retail money, it will be more susceptible to swings based on public sentiment. Therefore, we want to include the percent ownership attributed to institutional capital of the total float. We recognize that more data will be required to proceed with this analysis.

[7] https://www.jstor.org/stable/2325486

2. **Risk-Return Profile:** While public sentiment may be an important driver of value, we want to benchmark our analysis against more traditional daily indicators that could drive return, namely major market indices.

3. **Herd Effect:** Do returns seem more responsive to more popular positive or negative tweets, as investors follow popular advice?

4. **Past Feelings:** Do investors tend to consider past sentiment, or does sentiment fluctuate seemingly randomly each day?

4 Strategy

There are two specific areas we want to focus on in terms of answering the aforementioned research question and addressing our null hypothesis.

1. **Public Sentiment:** We want to use simple multi-linear regression to analyze the effect of positive and negative sentiment on daily returns. We focus on returns, rather than price, to see the intra-day swings in prices as investors internalize market sentiment. However, given the presence of confounding variables and the high propensity of noise, it is important to be methodical when predicting our dependent variables.

2. **Predicting Returns:** We want to build a machine learning model to predict daily returns. We want to explore different types of ML techniques and see how they perform along different measurements of model success (R-squared, ROC etc.). Ideally, we will analyze our specific stocks with models from sentiment scores and other indicators. In a perfect world, we will use a multi-layer feed forward neural network built in pytorch, but we will also experiment with a long-short-term memory neural network and other

models.

5 Data Sources & Collection

1. Twitter Data:

   (a) We will use a Tweet Collector, a package developed by Omer Metin on github[8] to scrape data from January 1, 2015 to December 31, 2019. We want to aggregate all the Tweet mentions of the selected asset's tickers.

   (b) We choose five companies to analyze: Amazon (AMZN), Apple (AAPL),Tesla (TSLA), Google (GOOG), and Microsoft (MSFT). We believe each of these companies is popular enough to generate a sufficient set of Tweets in our given time horizion; if we had chosen a smaller company, we may not find as much data. Moreover, tech companies are also more likely to generate mentions than companies in less B2C-focused industries. With these choices, we acknowledge that our analysis will only be generalizable to large-cap tech companies.

   (c) This data takes the form of a dataset with 7 features: Tweet ID, Writer, Post Date, Text, Number of Comments, Number of Retweets, and Number of Likes.

2. Wharton Research Data Services:

   (a) We pulled stock price data for AMZN, AAPL, TSLA, GOOG, and MSFT, as well as total market data, from January 1, 2015, to December 31, 2019 using WRDS. We selected data from a report by the Center for Research in Security Prices,

[8] https://github.com/omer-metin/TweetCollector

LLC.

(b) This data includes one stock data set with 8 features and one market data set with 5 features, which we will use as an additional predictor in our modeling stage.

6 Preprocessing and Data Engineering

Data will also have to be processed for sentiment — the text will have to be cleaned — we have to decide on a library to use. We decided on the VADER sentiment analysis package, found on github.[9] VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

We first filtered out Tweets with less than 10 likes, 5 comments, or 5 retweets. Given the large size of the data set of all tweets from 2015-2020, we reasoned that Tweets with so few interactions were not relevant to our analysis. Then, we filtered Tweets by specific company mentioned into five separate DataFrames. From here, we computed a sentiment score for each Tweet. We weighted our final "sentiment index" by comments according to the following formula:

$$sentiment_{index} = sentiment_{score} * (1.5 * comment_{num} + 1.25 * retweet_{num} + like_{num})$$

We reasoned that comments were the most meaningful form of engagement since they require the most effort on the part of the Twitter user, followed by retweets and finally likes. With this formula, we are able to normalize our sentiment score value to get a better proxy for how Twitter users are feeling about our selected companies.

[9] https://github.com/cjhutto/vaderSentiment

After computing a sentiment index value for each Tweet, we sorted Tweets by day and computed the mean Tweet sentiment for a given day. Then, we merged our grouped tweet data set with our stock data set to get a DataFrame of daily sentiment and stock info.

We also computed daily returns at each point in time for each stock according to the formula below because, as mentioned previously, we want to account for changes in price.

$Return = (P_{Close} - P_{Open})/P_{Open}$

7 EDA

7.1 Initial Analysis

After downloading, merging, wrangling, and cleaning our data, we ran summary statistics to ensure that the collected data seemed reasonable and was properly exported. There are 7,548 rows in our stock data set, while there are 3,717,964 Tweets in our tweet data set. The earliest and latest dates in both data sets are January 1, 2015 and December 31, 2019, respectively. We do not have any missing values, aside from 1,163 author-less Tweets. However, as the Tweet author is not relevant to our analysis, we proceed accordingly. There are no duplicate values in either data set.

After merging our data sets and separating them out by company, we are left with five separate data sets describing the price, volume, trades, shares outstanding for Apple, Amazon, Tesla, Google, and Microsoft for each day in our sample. No scaling is required for our data since all prices are quoted in USD and all sentiment scores are computed the same way. Each of these data sets has a number of missing values for sentiment index. Namely, there are many days in our sample time horizon where the company in question is simply not

Tweeted about. Below is a graph of each of the sentiment index series over time (Figure 1). Sentiment scores were not highly correlated (Figure 4).

Some companies seem to have larger spikes in sentiment scores than others, but it's clear that investors were very emotional over the course of our sample. Notably, we generally see large positive spikes in sentiment scores in the early months of the year, whereas we generally see large negative spikes in sentiment scores in the later months of the year. Given that stocks tend to do well in Q1 and worse in Q4, this plot is encouraging. The means and standard deviations for each company are remarkably similar, with means between 10 and 16 and standard deviations between 21 and 55. More Twitter users seem to think positively of our companies on any given day than those who seem to think negatively (Figures 2-3).

7.2 Primary EDA

There are notably a few outliers in our data sets, with a number of points falling outside each companies' sentiment indices' interquartile ranges. We choose not to remove these outliers, as they are points in our sample and there is no reason to suspect an error.

The returns of our selected companies do not appear to be highly correlated (see Figure 10). Moreover, we are not worried about high correlation between our other daily returns because we figure tech stock returns will be somewhat correlated.

We note that Tesla's returns seems to be significantly less correlated with other tech stocks we selected. Anecdotally, Tesla has been an extremely popular stock among retail traders due to Elon Musk's charismatic, magnetic personality. Additionally, Musk is an active Twitter user, often tweeting his thoughts about contemporary issues and Tesla initiatives. Moreover, the company isn't a 'pure' tech company like our other selected tech companies; Tesla is

primarily an electric vehicle and clean energy manufacturer, though they have explored other avenues in recent years.

The sentiment indices are not correlated with other predictors (see figures 4-9), Value-Weighted, Equal-Weighted, and S&P-500 Returns, but the other three variables are expectedly highly correlated because they contain baskets of similar securities. We are not worried about this high correlation, as we are bunching these in as external market indicators.

Sentiment scores and daily returns don't appear to be significantly related for our selected companies, observing Figure 22.

8 Modeling

8.1 Model Outcome and Features

8.1.1 Summary Stats

The summary statistics of our training dataframes can be found in the figures section (see figures 16-20). The daily returns of our selected companies look to be close, as do sentiment scores. Standard deviations are also similar.

8.1.2 Two Sample T-Tests

We conducted two sample T-tests for the training and testing data for each of our selected companies. Notably, the training and testing data are statistically significantly different for Google (Figures 10-15).

8.2 Model Building

8.2.1 Linear Naive Baseline

We break our model building up into two categories: features X and features Z. Features Z is our baseline analysis: we try and predict daily stock returns given daily sentiment. Features X is our benchmark analysis: we try to predict daily stock returns given daily sentiment and S&P500, Equal-Weighted, and Value-Weighted Returns.

We begin by fitting a simple linear model to the training data sets for each of our selected companies. The model scores (26.74% for Apple, 42.4% for Amazon, 8.1% for Tesla, 38.6% for Google, and 43.7% for Microsoft, an average of 32.0%) are not particularly high, and our mean CV scores are not significantly different (26.05% for Apple, 41.0% for Amazon, 7.5% for Tesla, 37.7% for Google, and 43.7% for Microsoft, an average of 41.2%). Notably, the model has low predictive power for Tesla, which is surprising; Tesla attracts many online followers, who we feel would be quick to Tweet their opinions. (See Linear Model Coefficients, Figures 23-24)

The linear model fit on Z scores significantly worse, losing the predictive power of the variables in X. Most of the scores are near zero, and we don't see significant improvement for cross-validated or tuned model scores.

The models score poorly on RMSEs. Performing a GridSearchCV for a linear fit, copy_X as True, fit_intercept as True, and normalize as False yield the best results. Still, we don't see a huge increase in scores across the board (26.74% for Apple, 42.4% for Amazon, 8.7% for Tesla, 38.9% for Google, and 43.74% for Microsoft)

8.3 Ridge, Lasso, and Elastic Net

We next fit a Ridge model with alpha = 0.5 on both X and Z. We actually see a decrease in scores, with 11.9% for Apple, 13.9% for Amazon, 3.9% for Tesla, 12.8% for Google, and 12.5% for Microsoft for X and values near zero for Z. The RMSE scores are low as well, and we also see low mean CV scores. A tuned Ridge model still does not improve scores significantly for X and Z, and it is beaten by the linear model. The ridge model performs worse with Z than with X. (See Ridge coefficients, Figures 25-26)

Fitting a Lasso model with alpha = 0.1, we see predictive performance similar to the linear model on the training data, for both the cross-validated and the tuned Grid Search model. Finally, we fit an elastic net model with alpha = 0.5. Again, we see predictive performance similar to the linear model, and again, the lasso model and elastic net model perform worse with Z than with X. (Figures 27-30)

8.3.1 Random Forest

We finish with a tree-based model: a Random Forest regression. The model scores very highly for random forests: Apple 88.0%, Amazon 91.0%, Tesla 85.0%, Google 90.0%, and Microsoft 90.0%. Nonetheless, our mean CV scores and tuned random forest model performs more poorly: Apple 21.5%, Amazon 27.3%, Tesla 4.09%, Google 27.3%, and Microsoft 28.7%.

The random forest model still performs well using the Z_train data. As a result, it is our preferred model, and we choose to use it on the testing data. However, given the relative difficulty in regards to interpretability of more complex learning algorithims, we will focus on analyzing the results of the model in terms of our earlier hypotheses as well as tie the model back to some of the behavioral factors explored earlier.

Given that we were analyzing sentiment, an important assumption we made was that in regards to market reactivity. We assumed that daily sentiment would have an almost instantaneous effect on trading activity within the same day. However, in reality, there might be lags associated with positive sentiment and market reactions. For example, since earnings news comes out after the market has closed, positive sentiment on that day (if earnings beat expectations) could have minimal effect on that day trading activity, but a out sized effect on the next day.

8.4 Predictive Performance on Testing Data

Now, we fit our Random Forest model on the test data. Our model performs poorly on X and Z, and the resulting R-squared values are shown in Figure 31 fit with the parameters in Figure 21.

From an initial analysis of the mean squared error of our model for each of our selected companies, it appears even the random forest model performs poorly on the test set. As such, we would not recommend this model to a client seeking to understand how daily returns are predicted.

9 Conclusion

While we found market sentiment could indeed reliably predict daily returns for Apple, Amazon, Tesla, and Google on the training data, the poor predictability of our chosen model on testing data implies there are other complexities in asset pricing that can better account for price movements. Factors that we originally expected to draw on the relationship, such as volume composition, risk-return profile, the herd effect, and past feelings all also likely

play a role. Nonetheless, our initial hypothesis that market sentiment for stocks would be highly correlated with daily returns looks to be somewhat true and at least worth further exploration. We ran several models to explore the data we collected from Twitter and WRDS, and found the scores from our Random Forest regression to be sufficient, and significantly higher than our output from other models such as Linear, Ridge, Lasso, and Elastic Net.

9.1 Next Steps

Our analysis was restricted to large tech companies, and it would be interesting to see if the results hold for other types of companies. Moreover, our results may not be generalizable to all S&P 500 companies. We could also explore other regression models, including but not limited to: stochastic gradient descent, decision trees, support vector regression, polynomial regression, quantile regression, step-wise regression, or partial least squares regression. These could all be potential avenues of further exploration.

There is widely-available literature for further reading on the relationship between investor sentiment and daily returns. For example, Overnight Returns and Firm-Specific Investor Sentiment by David Aboody, Omri Even-Tov, Reuven Lehavy, and Brett Trueman supports using overnight returns to measure firm-specific sentiment. Moreover, A dynamic analysis of the relationship between investor sentiment and stock market realized volatility: Evidence from China by Yanhui Chen, Hanhui Zhao, Ziyu Li, and Jinrong Lu shows that investor sentiment indeed forecasts the realized volatility. Finally, The Effect of Online Investor Sentiment on Stock Movements: An LSTM Approach by Gaoshan Wang, Guangjin Yu, and Xiaohong Shen shows that an increase in investor sentiment can boost the major money flows in the market to some extent.

10 References

1. Aboody, D., Even-Tov, O., Lehavy, R., Trueman, B. (2018). Overnight returns and firm-specific investor sentiment. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2554010

2. Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science, [online] 2(1), pp.1-8. Available at: ¡https://arxiv.org/abs/1010.3003¿ [Accessed 2 December 2021].

3. Chen, S. T., Haga, K. Y. (2021). Using E-GARCH to analyze the impact of investor sentiment on stock returns near stock market crashes. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.664849

4. Chen, Y., Zhao, H., Li, Z., Lu, J. (2020). A dynamic analysis of the relationship between investor sentiment and stock market realized volatility: Evidence from China. PLOS ONE, 15(12). https://doi.org/10.1371/journal.pone.0243080

5. C.J. Hutto, vaderSentiment, (2021), GitHub repository, https://github.com/cjhutto/vaderSentiment

6. Dobbs, R., Huyett, B. and Koller, T., 2021. The CEO's guide to corporate finance. [online] mckinsey.com. Available at: ¡https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-ceos-guide-to-corporate-finance¿ [Accessed 2 December 2021].

7. Fama, E., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, [online] 25(2). Available at: ¡https://www.jstor.org/stable/2325486¿ [Accessed 2 December 2021].

8. Leask, H., 2021. Hedge funds' use of alternative data tipped to surge, new industry study finds. [online] Hedgeweek. Available at: ¡https://www.hedgeweek.com/2020/05/04/285283/he funds-use-alternative-data-tipped-surge-new-industry-study-finds¿ [Accessed 2 December 2021].

9. Moldovan, E., 2021. Investors Psychology And The Herd Effect On The Financial Markets. [online] Ideas.repec.org. Available at: ¡https://ideas.repec.org/a/aio/rteyej/v1y2010i15sp21-26.html¿ [Accessed 2 December 2021].

10. Omer Metin, TweetCollector, (2020), GitHub repository, https://github.com/omer-metin/TweetCollector

11. Perrin, A. and Atske, S., 2021. About three-in-ten U.S. adults say they are 'almost constantly' online. [online] Pew Research Center. Available at: ¡https://www.pewresearch.org/fact-tank/2021/03/26/about-three-in-ten-u-s-adults-say-they-are-almost-constantly-online/¿ [Accessed 2 December 2021].

12. Seok, S. I., Cho, H., Ryu, D. (2019). Firm-specific investor sentiment and daily stock returns. The North American Journal of Economics and Finance, 50, 100857. https://doi.org/10.1016/j.najef.2018.10.005

13. Statista. 2021. U.S. Twitter reach by age group 2021 — Statista. [online] Available at: ¡https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/¿ [Accessed 2 December 2021].

14. Wang, G., Yu, G., Shen, X. (2020). The effect of online investor sentiment on Stock Movements: An LSTM approach. Complexity, 2020, 1–11. https://doi.org/10.1155/2020/4754025

11 Appendix

11.1 Figures

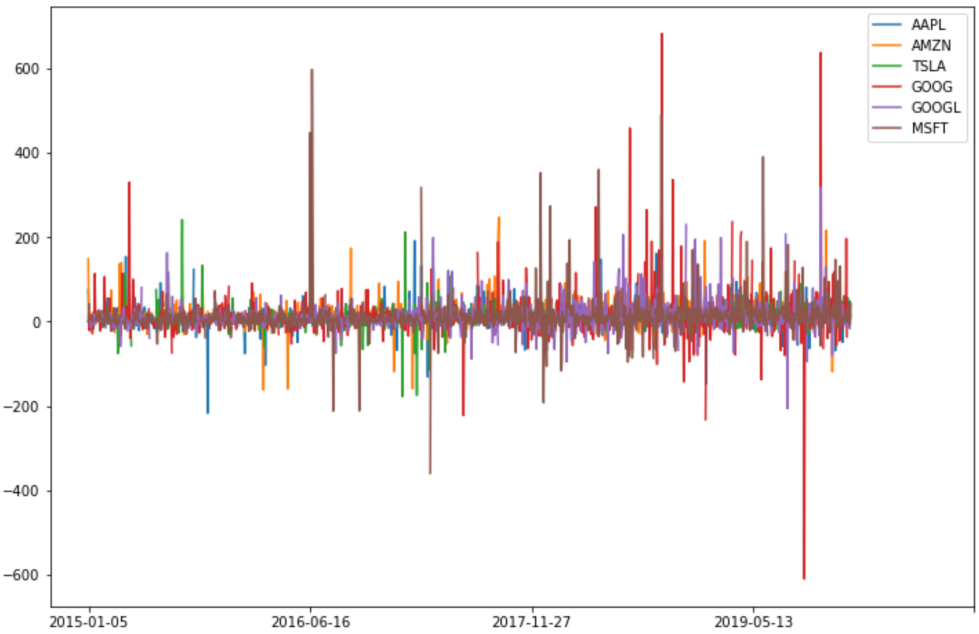

Figure 1: Series of Sentiment Scores from 2015-2020

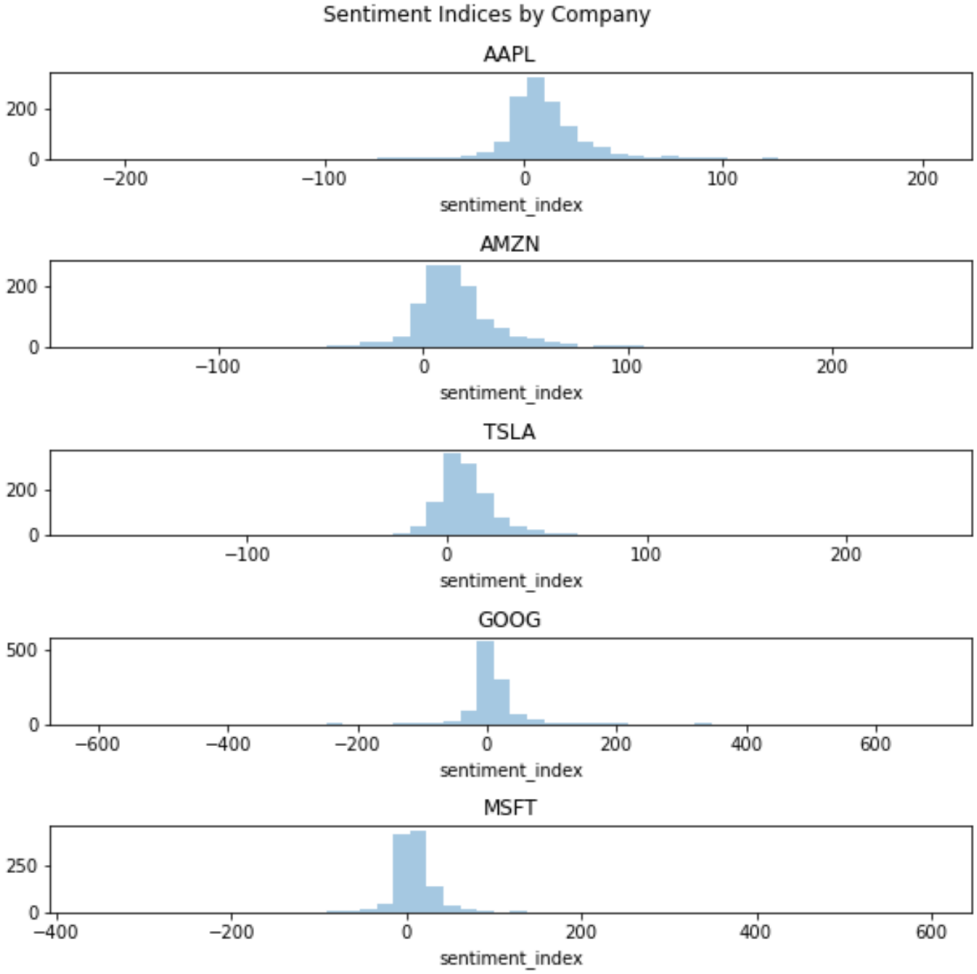

Figure 2: Distributions of Sentiment Scores

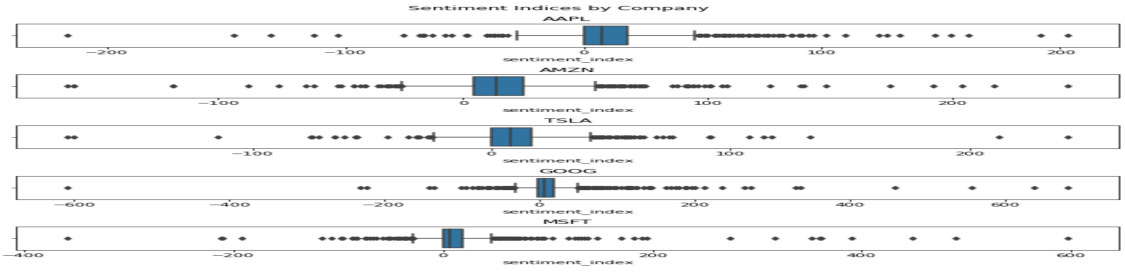

Figure 3: Boxplots of Sentiment Scores

	AAPL	AMZN	TSLA	GOOG	MSFT
AAPL	1.000000	0.176234	0.126418	0.061245	0.153431
AMZN	0.176234	1.000000	0.060031	0.165126	0.090680
TSLA	0.126418	0.060031	1.000000	-0.002128	0.038054
GOOG	0.061245	0.165126	-0.002128	1.000000	0.035478
MSFT	0.153431	0.090680	0.038054	0.035478	1.000000

Figure 4: Correlation between Sentiment Scores

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment	daily_return
Value_Weighted_Return_Dist	1.000000	0.972720	0.999861	0.100206	0.512194
Equal_Weighted_Return_Dist	0.972720	1.000000	0.973200	0.083686	0.454129
SP500_Return	0.999861	0.973200	1.000000	0.099576	0.511512
sentiment	0.100206	0.083686	0.099576	1.000000	0.116405
daily_return	0.512194	0.454129	0.511512	0.116405	1.000000

Figure 5: Correlation between Daily Predictors for AAPL

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment	daily_return
Value_Weighted_Return_Dist	1.000000	0.972752	0.999861	0.032449	0.530551
Equal_Weighted_Return_Dist	0.972752	1.000000	0.973237	0.034597	0.450007
SP500_Return	0.999861	0.973237	1.000000	0.032875	0.526329
sentiment	0.032449	0.034597	0.032875	1.000000	0.023427
daily_return	0.530551	0.450007	0.526329	0.023427	1.000000

Figure 6: Correlation between Daily Predictors for AMZN

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment
Value_Weighted_Return_Dist	1.000000	0.972678	0.999861	0.001180
Equal_Weighted_Return_Dist	0.972678	1.000000	0.973158	0.000754
SP500_Return	0.999861	0.973158	1.000000	0.001404
sentiment	0.001180	0.000754	0.001404	1.000000

Figure 7: Correlation between Daily Predictors for TSLA

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment
Value_Weighted_Return_Dist	1.000000	0.972897	0.999857	-0.020366
Equal_Weighted_Return_Dist	0.972897	1.000000	0.973398	-0.016422
SP500_Return	0.999857	0.973398	1.000000	-0.020826
sentiment	-0.020366	-0.016422	-0.020826	1.000000

Figure 8: Correlation between Daily Predictors for GOOG

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment
Value_Weighted_Return_Dist	1.000000	0.971881	0.999862	0.022365
Equal_Weighted_Return_Dist	0.971881	1.000000	0.972403	0.013707
SP500_Return	0.999862	0.972403	1.000000	0.022644
sentiment	0.022365	0.013707	0.022644	1.000000

Figure 9: Correlation between Daily Predictors for MSFT

	AAPL	AMZN	TSLA	GOOG	MSFT
AAPL	1.000000	0.560770	0.339419	0.577407	0.597573
AMZN	0.560770	1.000000	0.378765	0.733023	0.637437
TSLA	0.339419	0.378765	1.000000	0.371527	0.339567
GOOG	0.577407	0.733023	0.371527	1.000000	0.672002
MSFT	0.597573	0.637437	0.339567	0.672002	1.000000

Figure 10: Correlation between Returns for Selected Companies

	tr_avg	te_avg	tdiff
Value_Weighted_Return_Dist	0.000576	0.000018	0.902767
Equal_Weighted_Return_Dist	0.000479	0.000109	0.592168
SP500_Return	0.000494	-0.000062	0.900101
sentiment	11.073540	10.116257	0.463730
daily_return	0.000753	-0.000425	1.373501

Figure 11: Two Sample T-Test AAPL

	tr_avg	te_avg	tdiff
Value_Weighted_Return_Dist	0.000390	0.000962	-1.045858
Equal_Weighted_Return_Dist	0.000357	0.000775	-0.755021
SP500_Return	0.000313	0.000868	-1.015842
sentiment	16.399362	16.827414	-0.241829
daily_return	-0.000667	0.002403	-3.161865

Figure 12: Two Sample T-Test AMZN

	tr_avg	te_avg	tdiff
Value_Weighted_Return_Dist	0.000655	0.000056	1.057037
Equal_Weighted_Return_Dist	0.000626	-0.000098	1.243233
SP500_Return	0.000573	-0.000020	1.045216
sentiment	9.263907	11.043583	-1.226776
daily_return	0.000869	-0.000758	1.071349

Figure 13: Two Sample T-Test TSLA

	tr_avg	te_avg	tdiff
Value_Weighted_Return_Dist	0.000786	-0.000348	1.707720
Equal_Weighted_Return_Dist	0.000763	-0.000467	1.833537
SP500_Return	0.000703	-0.000424	1.697035
sentiment	12.580316	12.652251	-0.021007
daily_return	0.000673	-0.000849	1.770369

Figure 14: Two Sample T-Test GOOG

	tr_avg	te_avg	tdiff
Value_Weighted_Return_Dist	0.000545	0.000319	0.348433
Equal_Weighted_Return_Dist	0.000495	0.000238	0.390793
SP500_Return	0.000462	0.000251	0.326070
sentiment	11.671013	14.106071	-0.678807
daily_return	0.000046	-0.000542	0.542685

Figure 15: Two Sample T-Test MSFT

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment	daily_return
count	998.000000	998.000000	998.000000	998.000000	998.000000
mean	0.000576	0.000479	0.000494	11.073540	0.000753
std	0.008366	0.008540	0.008372	23.118574	0.012003
min	-0.040869	-0.041088	-0.040979	-216.750000	-0.066331
25%	-0.002815	-0.003444	-0.002855	0.147135	-0.005343
50%	0.000540	0.000544	0.000525	7.948052	0.000988
75%	0.004854	0.005365	0.004687	18.993056	0.007445
max	0.039178	0.033484	0.039034	191.541667	0.086961

Figure 16: Parameters of Random Forest Regression, Test Data

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment	daily_return
count	985.000000	985.000000	985.000000	985.000000	985.000000
mean	0.000390	0.000357	0.000313	16.399362	-0.000667
std	0.008737	0.008908	0.008741	27.382499	0.014536
min	-0.040869	-0.041088	-0.040979	-161.875000	-0.085615
25%	-0.002928	-0.003534	-0.002956	4.795455	-0.007359
50%	0.000547	0.000564	0.000509	13.609375	-0.000253
75%	0.004847	0.005338	0.004685	24.535714	0.007382
max	0.049762	0.045981	0.049594	247.295455	0.074520

Figure 17: Parameters of Random Forest Regression, Test Data

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment	daily_return
count	992.000000	992.000000	992.000000	992.000000	992.000000
mean	0.000655	0.000626	0.000573	9.263907	0.000869
std	0.008627	0.008748	0.008628	21.307731	0.023519
min	-0.040869	-0.041088	-0.040979	-177.673077	-0.081792
25%	-0.002595	-0.003111	-0.002652	0.222529	-0.012994
50%	0.000528	0.000569	0.000498	8.080357	0.000752
75%	0.004818	0.005252	0.004678	16.253348	0.013152
max	0.049762	0.045981	0.049594	241.458333	0.135137

Figure 18: Parameters of Random Forest Regression, Test Data

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment	daily_return
count	906.000000	906.000000	906.000000	906.000000	906.000000
mean	0.000786	0.000763	0.000703	12.580316	0.000673
std	0.008350	0.008509	0.008354	58.136121	0.012147
min	-0.040869	-0.041088	-0.040979	-609.500000	-0.057637
25%	-0.002487	-0.003017	-0.002614	-1.492188	-0.005503
50%	0.000642	0.000698	0.000588	5.862500	0.000643
75%	0.004881	0.005377	0.004734	18.869792	0.007810
max	0.049762	0.045981	0.049594	682.250000	0.052006

Figure 19: Parameters of Random Forest Regression, Test Data

	Value_Weighted_Return_Dist	Equal_Weighted_Return_Dist	SP500_Return	sentiment	daily_return
count	921.000000	921.000000	921.000000	921.000000	921.000000
mean	0.000545	0.000495	0.000462	11.671013	0.000046
std	0.008268	0.008421	0.008278	45.530510	0.014204
min	-0.039372	-0.041088	-0.039414	-211.625000	-0.072964
25%	-0.002827	-0.003436	-0.002913	0.000000	-0.007288
50%	0.000503	0.000547	0.000489	6.875000	0.000093
75%	0.004787	0.005338	0.004660	18.250000	0.007589
max	0.049762	0.045981	0.049594	597.295455	0.074520

Figure 20: Parameters of Random Forest Regression, Test Data

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'mse',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 24,
 'verbose': 0,
 'warm_start': False}
```

Figure 21: Parameters of Random Forest Regression, Test Data

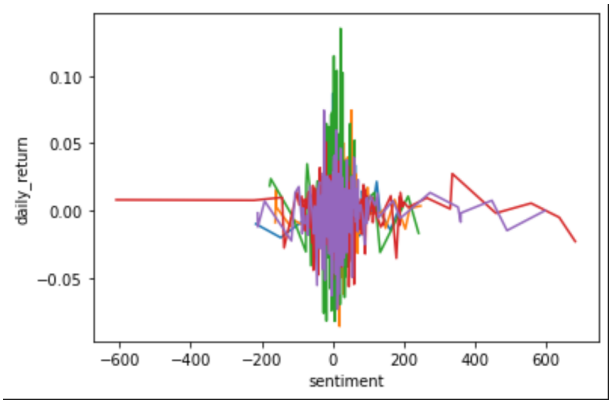

Figure 22: Sentiment scores vs. Daily Returns

```
Best Linear Model Coefficients: [ 2.1157 -1.0082 -0.4256  0.      0.    ]
Best Linear Model Coefficients: [ 26.0443  -1.7896 -23.3629   0.      0.    ]
Best Linear Model Coefficients: [ 11.6158  -0.0236 -10.8385   0.      0.    ]
Best Linear Model Coefficients: [11.4389 -1.619  -9.0456 -0.      0.    ]
Best Linear Model Coefficients: [ 24.1669  -1.97   -21.2652  -0.      0.    ]
```

Figure 23: Coefficients, Linear Model X

```
Best Linear Model Coefficients: [0.0001 0.    ]
Best Linear Model Coefficients: [0. 0.]
Best Linear Model Coefficients: [0. 0.]
Best Linear Model Coefficients: [-0.  0.]
Best Linear Model Coefficients: [-0.  0.]
```

Figure 24: Coefficients, Linear Model Z

```
Best Ridge Model Coefficients: [ 2.1157 -1.0082 -0.4256  0.      0.    ]
Best Ridge Model Coefficients: [ 26.0443  -1.7896 -23.3629   0.      0.    ]
Best Ridge Model Coefficients: [ 11.6158  -0.0236 -10.8385   0.      0.    ]
Best Ridge Model Coefficients: [11.4389 -1.619  -9.0456 -0.      0.    ]
Best Ridge Model Coefficients: [ 24.1669  -1.97   -21.2652  -0.      0.    ]
```

Figure 25: Coefficients, Ridge Model X

```
Best Ridge Model Coefficients: [0.0001 0.    ]
Best Ridge Model Coefficients: [0. 0.]
Best Ridge Model Coefficients: [0. 0.]
Best Ridge Model Coefficients: [-0.  0.]
Best Ridge Model Coefficients: [-0.  0.]
```

Figure 26: Coefficients, Ridge Model Z

```
Best Lasso Model Coefficients: [0. 0. 0. 0. 0.]
Best Lasso Model Coefficients: [0. 0. 0. 0. 0.]
Best Lasso Model Coefficients: [0. 0. 0. 0. 0.]
Best Lasso Model Coefficients: [ 0.  0.  0. -0.  0.]
Best Lasso Model Coefficients: [ 0.  0.  0. -0.  0.]
```

Figure 27: Coefficients, Lasso Model X

```
Best Elastic Net Model Coefficients: [0. 0. 0. 0. 0.]
Best Elastic Net Model Coefficients: [0. 0. 0. 0. 0.]
Best Elastic Net Model Coefficients: [0. 0. 0. 0. 0.]
Best Elastic Net Model Coefficients: [ 0.  0.  0. -0.  0.]
Best Elastic Net Model Coefficients: [ 0.  0.  0. -0.  0.]
```

Figure 28: Coefficients, Elastic Net Model Z

```
Best Lasso Model Coefficients: [0. 0.]
Best Lasso Model Coefficients: [0. 0.]
Best Lasso Model Coefficients: [0. 0.]
Best Lasso Model Coefficients: [-0.  0.]
Best Lasso Model Coefficients: [-0.  0.]
```

Figure 29: Coefficients, Lasso Model X

```
Best Elastic Net Model Coefficients: [0. 0.]
Best Elastic Net Model Coefficients: [0. 0.]
Best Elastic Net Model Coefficients: [0. 0.]
Best Elastic Net Model Coefficients: [-0.  0.]
Best Elastic Net Model Coefficients: [-0.  0.]
```

Figure 30: Coefficients, Elastic Net Model Z

```
R^2 Value: 0.00022953008472337145
R^2 Value: 0.00024369720146255976
R^2 Value: 0.0006182129580928295
R^2 Value: 0.0001792393280275524
R^2 Value: 0.00028898259934173034


R^2 Value: 9.901692947567013e-05
R^2 Value: 0.00012818495168285347
R^2 Value: 0.00046909476032926096
R^2 Value: 7.961179460809819e-05
R^2 Value: 0.00018893304493317677
```

Figure 31: R-squared Scores for Random Forest Models, Test Data