Thesis

Liam Connor Murray
Karl Olof Vincent Lindberg

April 2023

Contents

Abstract

Our abstract revolves around the failure of socialism thrpughput history

1 Introduction

1.1 Background

There exists a number of definitions for investor sentiment that can be found in behavioral finance literature, although the vast majority of them refer to investor sentiment at an aggregate, market-wide level and not at an individual, firm-specific level. Authors Malcom Baker and Jeffrey Wurlger offer up two definitions of market wide sentiment in their frequently cited journal article titled "Investor Sentiment and the Cross-Section of Stock Returns" published in 2004. The first definition of investor sentiment refers to the propensity of an investor or group of investors to speculate, while the second definition refers to an investor or group of investors' optimism or pessimism towards stocks in general. Seok et al define investor sentiment as the resulting demand shocks generated by uninformed, noisy investors that lead to persistent mispricings in asset prices (Seok et al, 2019). In the theoretical study titled "Market liquidity as a sentiment indicator", authors Malcom Baker and Jeremy C. Stein identifies "sentiment shocks" occurring as a result of overconfident, uninformed investors over weighting the strength of their private signals (either positive buying signals or negative selling signals), and then (perhaps irrationally) trading on these signals accordingly.

Despite the existence of the ever-growing literature demonstrating that excess returns on stocks are not completely explained by their respective fundamentals, there exists no clear-cut, unanimously agreed upon method of measuring and/or quantifying investor sentiment. Authors Baker and Wurgler state that "there are no definitive or uncontroversial measures" with regards to investor sentiment (Baker and Wurgler, 2004). In the realm of factor investing, there exists a well-known and often cited framework known as the Fama-French three factor model often used to explain future expected returns. The Fama-French three factor model includes market risk; measured by the market return in excess of the risk free rate, size; measured by the difference in returns of smaller capitalization stocks and large capitalization stocks, and value; measuring the difference in returns of firms with high book-to-market ratios (referred to as value stocks) and firms with low book-to-market stocks (referred to as growth stocks). In the paper titled "Measures Of Investor Sentiment: A Comparative Analysis Put-Call Ratio Vs. Volatility Index" published in August 2008 by the Journal of Business Economic Research, authors Jones and Bandopadhyaya state that "non-economic factors such as investor sentiment are increasingly being recognized as explanatory variables for analyzing asset prices" and "as the literature grows, so too does the array of competing measures" (Bandopadhyaya and Jones, 2008).

Fast forward to today at the present time of writing, there is still no well-documented, unanimously agreed upon framework to quantify or measure investor sentiment, at either a market-wide or firm-specific level.

1.2 Investor Sentiment Vs. Investor Attention

Investor attention and investor sentiment are two distinct concepts in finance that can provide valuable insights into the behavior, tendencies, and general attitudes of market participants. Investor attention refers to the level of awareness that investors possess towards a specific stock, sector/industry or market as a whole. Investor attention can be measured and quantified through a variety of different metrics such as the volume of trades, the level of media coverage, the number of research reports or the aggregate count of analyst buy/sell recommendations for a specific stock, industry or market. Investor sentiment on the other hand refers to the general mood or emotion investors and market participants convey towards current market conditions. Investor sentiment is indicative of the collective level of confidence (skepticism) and optimism (pessimism) that market participants currently feel towards market expectations at some arbitrary point in the future. The important distinguishing factor between investor sentiment and investor attention is the dimensionality of each concept. Investor sentiment is 3 dimensional, and can be quantifiable as relatively positive (bullish), relatively negative (bearish), or relatively neutral. Investor attention on the other hand is two dimensional and is therefore measurable on a single axis from relatively high to relatively low.

Kahneman, Daniel., 1973. Attention and Effort (Prentice-Hall, Englewood Cliffs, NJ).

In addition to the differences in dimensionality, investor attention must lead investor sentiment. For an investor to feel either bearish or bullish about a specific stock, sector/industry or market, an investor must first be aware and have allocated a portion of their attention accordingly prior to having the capacity of developing some degree of sentiment. Investor attention is limited and scarce in nature (Kahneman, 1973), and investors must determine how to allocate their attention across the tens of thousands of listed stocks worldwide (World Federation of Exchanges, 2021). As mentioned in the Journal of Finance article titled "In Search of Attention" by authors Da, Engelberg and Gao published in 2011, there does not exist a direct proxy of investor attention. It is virtually impossible to measure and/or quantify what fraction of an individual's mental capacity is consumed by a specific stock at any given point in time. Instead, a variety of indirect proxies are used to relay an understanding of the allocational degree of investor attention

toward a specific stock, sector/industry or market.

Indirect proxies of investor attention include but are not limited to trading volume, market turnover rate, frequency of mention on social media platforms (i.e. Twitter, Reddit etc.), frequency of mention in financial news articles and/or headlines (i.e. Dow Jones, Wall Street Journal, Reuters etc.) and online search volume frequency (i.e. Google Trends). It is important to note that the sole mention of a stock, sector/industry or market in prominent financial news media does not guarantee an allocation of investor attention, which is "especially true in the so-called information age where a wealth of information creates a poverty of attention" (Da et al, 2011). Trading volume and market turnover can fluctuate for a plethora of different reasons besides investor attention, and "a news article in the Wall Street Journal does not guarantee attention unless investors actually read it" (Da et al, 2011). If all investors who stumble upon a financial news article concerning a stock enter a trade upon finishing reading, the aggregate count of published financial news articles in a specific period of time would be a strong quantifier of investor attention. Perhaps all investors prefer to search a stock's ticker online prior to making a trading decision. Investors may want to see recent price development or confirm if the tone of other financial news outlets are aligned with the general tone of the initial article. In this case, the aggregate count of a stock's ticker as a search term may serve as a better proxy for investor attention.

1.3 Investor Sentiment Proxies

`https://www.cboe.com/tradable_products/vix/`

There exists today a number of investor sentiment gauges at a market-wide level with the most frequently cited being the VIX. The VIX, or more formally the CBOE Volatility Index, is an index that measures the expected volatility of the SP 500 index over the following 30 days. Financial news media frequently refers to the VIX as the "fear index" as it tends to rise during periods of heightened uncertainty and fall during periods of market stability and strengthening investor confidence. As stated on the CBOE's Vix Volatility Suite web page, "the VIX index is a calculation designed to produce a measure of constant 30-day expected volatility of the U.S. stock market, derived from real-time, mid-quote prices of SP 500 Index (SPX) call and put options" (Chicago Board Options Exchange, 2023). In addition to the VIX, a very similar indicator known as the VXN provides a similar insight, instead using the Nasdaq-100 index as the underlying. The Nasdaq-100 is heavily weighted towards technology, and growth oriented stocks, which tend to be more difficult to value, more difficult to arbitrage, and as a result, more

susceptible to sentiment-based price movement (Baker and Wurgler, 2004). For a technology focused investor, the VXN is a more appropriate, sector specific gauge of the 30-day implied volatility of their chosen market.

Zhang, Q.T., Li, B., Xie, D. (2022). Sentiment Factors in Finance. In: Alternative Data and Artificial Intelligence Techniques. Palgrave Studies in Risk and Insurance. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-11612-4˙9

In addition to using the implied volatility of 30-day index options, investor sentiment can be imperfectly measured using a variety of other indicators such as the bullish percent index, high-low index and put-call ratios. The bullish percent index, often referred to as the BPI, is a market breadth indicator commonly used by market technicians to gauge the overall health of the market. The intuition behind the BPI is fairly straightforward, albeit rather rudimentary. The BPI is calculated by identifying the number of stocks forming point-and-figure and bullish reversal patterns, and simply dividing the number of stocks in a bullish trend by the total number of stocks in the index or sector of interest (Zhang, Q.T., Li, B., Xie, D., 2022). A BPI value ranging from 0-49% is indicative of negative market sentiment. A BPI value ranging from 51-100% is indicative of positive market sentiment. Lastly, a BPI value of exactly 50% is indicative of neutral market sentiment.

The high-low index (HLI) is another frequently used technical analysis indicator that intends to provide a gauge of investor sentiment. The HLI records the difference between the highest high and lowest low of a stock or market index over a specified look back period, ultimately measuring the strength or weakness of a trend or trend reversal (Zhang, Q.T., Li, B., Xie, D., 2022). Calculating the high-low index involves subtracting the lowest low from the highest high, and dividing the difference by the sum of the lowest low and the highest high to generate a value ranging from -100 to 100. Positive values indicate an uptrend in the underlying, while negative values indicate a downtrend in the underlying. Although indicators such as the BPI or HLI intend to gauge investor sentiment, by using stock price data, these indicators provide an indirect proxy for investor sentiment. Indicators using price data measure the second order effect of investor sentiment on a stock's price as opposed to providing a direct measurement of investor sentiment.

Bandopadhyaya, A., Jones, A. L. (2008). Measures Of Investor Sentiment: A Comparative Analysis Put-Call Ratio Vs. Volatility Index. Journal of Business Economics Research (JBER), 6(8). https://doi.org/10.19030/jber.v6i8.2458

A more direct proxy for investor sentiment is the put-call ratio. The put-call ratio is calculated by dividing the total open interest in put options by the total open interest in call options for a specific stock. The put-call ratio is often referred to by practitioners when referring to general market

sentiment and is frequently used as a contrarian indicator (Bandopadhyaya and Jones, 2008). Authors Bandopadhyaya and Jones conclude that in the period of January 2004 through April 2006, the put-call ratio is a statistically significant determinant of the value of the SP not explained by the previous day's value. An American put option gives an investor the right, but not the obligation, to sell an underlying asset at a specific price (the strike price) within a specified time period. An American call option gives an investor the right, but not the obligation, to buy an underlying asset at a specific price (the strike price) within a specified time period. A put-call ratio equal to 1 indicates that investors currently hold an equal number of open positions in both a stock's put and call options. A ratio greater (less) than 1 indicates investors currently hold more open positions in put (call) options, suggesting a relatively bearish (bullish) sentiment within the options market for a specific stock. A put-call ratio can provide a more direct, first order measurement of investor sentiment since price data is not used in its calculation and as a result may reflect aggregate investor sentiment in a more timely manner than price-based sentiment indicators such as the BPI or HLI outlined previously.

1.4 US Market Environment

Sharma, R. (2020). The Comeback Nation. Foreign Affairs. https://fred.stlouisfed.org/series/FEI

In the second half of 2009 shortly after the tail end of the global financial crisis of 2008, most major stock market indices worldwide began recovering their losses. Many governments and central banks around the world took significant measures to stabilize the global economy and stimulate economic growth. The United States emerged from the global financial crisis stronger than ever as "the U.S. stock market rose by 250 percent in the 2010s, nearly four times the average gain in other national stock markets" (Ruchir Sharma, 2020). For over an entire decade, U.S. equity markets enjoyed a utopian-like bull market primarily due to the low-interest rate environment. From December of 2008 until November 2015, the United States Federal Reserve's effective rate was held close to 0% (Board of Governors of the Federal Reserve System, 2023).

https://www.washingtonpost.com/business/economy/as-2010s-conclude-investors-have-enjoyed-bull-market-for-the-ages–but-many-americans-have-been-left-out/2019/12/31/da76a8a0-282e-11ea-ad73-2fd294520e97˙story.html

In addition to a prolonged low interest rate environment, quantitative easing (QE) measures by the United States Federal Reserve also contributed significantly to the steady, unwavering economic growth of the 2010s. QE is one of the many tools used by a central bank to conduct monetary policy and stimulate a slowing economy. QE refers to central banks purchasing

large amounts of financial assets such as government bonds from commercial banks and other financial institutions, effectively increasing the amount of money in circulation and providing substantial amounts of liquidity to financial markets. Additionally, increased amount of money in circulation has a downward effect on interest rates rendering it easier for individuals and businesses to borrow money to reinvest in the economy. As a result of the favorable economic conditions in the 2010s, the SP 500 (often used as the bench market index for the U.S economy) averaged an annual return of 13.5% (Washington Post, 2019).

1.5 Covid-19 Pandemic

1.6 FAANG+M

1.7 Bloomberg Social Velocity

Bloomberg LP (2014, February 20). Embedded Value in Bloomberg News and Social Sentiment Data, Sentiment Analysis White Paper. https://www.Bloomberg.com/. Retrieved January 5, 2023, from https://www.Bloomberg.com/professional/sentiment-analysis-white-papers/

In late 2014, Bloomberg added social sentiment analytics to their popular and widely used Bloomberg trading terminals. Within the Bloomberg terminal itself, the social sentiment analytics data is referred to as "Bloomberg Social Velocity". In the Bloomberg Professional Service Offering sentiment analysis white paper titled "Embedded Value In Bloomberg News Social Sentiment Data", the author begins by echoing the findings of previous sentiment-focused literature reiterating that "when rational arbitrageurs have limited risk-bearing capacity and time horizons, the actions of irrational noise traders can affect asset prices" and subsequently, "such actions can be interpreted as being driven by fluctuating investor sentiment" (Bloomberg, 2014). Through the use of supervised machine-learning techniques to process mass amounts of textual information, Bloomberg constructs both news and social sentiment values for a given ticker. It is important to note that social sentiment values are limited to the social media platform Twitter.

Based on the methodology outlined in the white paper, a human expert initially assigns scores (1 for positive sentiment , 0 for neutral sentiment, -1 for negative sentiment) to a financial news article or tweet. The score labeling is explicitly based on the question "if an investor having a long position in the security mentioned were to read this news or tweet, is he/she bullish, bearish or neutral on his/her holdings?" (Bloomberg, 2014). The annotated scores are then fed into machine-learning models, whereby the model automatically

assigns a confidence interval that a news article or tweet exhibits positive, neutral or negative sentiment concerning a specific stock.

`https://blog.twitter.com/official/en_us/a/2011/numbers.html`

The sentiment analytics data consists of six different observations including Twitter Publication Count (TC), Twitter Positive Count (TPC), Twitter Negative Count (TNC), News Publication Count (NC), News Positive Count (NPC) and News Negative Count (NNC). According to the white paper methodology, news sentiment data is "recomputed every two minutes with an eight hour rolling window" while Twitter sentiment data is "recomputed every minute with a 30-minute rolling window". It can be assumed that the recomputing window for Twitter sentiment data is half of that of news sentiment data since the volume of published tweets is far greater than the volume of financial news published in any arbitrary time period. In an official blog post published on March 14, 2011, Twitter revealed that the average number of tweets sent per day was 50 million (Twitter, 2011). Fast forwarding to the time of writing in 2023, the average number of tweets sent per day is likely far greater.

The Bloomberg news and social sentiment data paper also explores a number of trading strategies using sentiment analytics including a daily sentiment-strategy, a daily earnings event-driven strategy and an intraday sentiment-driven strategy, which are effectively confirming the efficient market hypothesis (FAMA-FRENCH REFERENCE). According to the back tested results, Bloomberg concludes that the "sentiment strategies outperform the corresponding benchmark index ETFs significantly, which strongly demonstrates the value embedded in Bloomberg News Social Sentiment data" (Bloomberg, 2014). Bloomberg reports a variety of metrics such as annualized return, annualized volatility, Sharpe ratio and average number of long and short positions. The daily earnings event-driven strategy was back tested for SP 500 stocks, Russell 3000 stocks and Russell 2000 stocks in order of descending average market capitalization throughout the period of January 2, 2015 to August 31, 2016. Based on the results of these backtests, Bloomberg concludes that the proposed trading strategies based on the Bloomberg News Social Sentiment data is optimal for SP 500 stocks because "SP 500 companies attract more attention and analyst coverage, so their average sentiment from news and social sources just before earnings are reported is more likely to contain earnings-related information" (Bloomberg, 2014). In addition to these findings, Bloomberg also reports that the highest Dharpe ratios are achieved between 5 minute and 30 minute trading intervals, reporting that beyond 30 minute trading intervals, hardly any meaningful returns are to be earned, ultimately reflecting the efficiency of financial markets with regards to incorporating new information (Bloomberg, 2014).

2 Literature Review

As our paper tries to explore and explain the significance of attention and sentiment when it comes to predicting stock returns our focus when reviewing literature will lie on sentiment and attention in the stock market.

2.1 Investor Sentiment

2.1.1 Measures Of Investor Sentiment: A Comparative Analysis Put-Call Ratio Vs. Volatility Index

Bandopadhyaya, A., Jones, A. L. (2008). Measures Of Investor Sentiment: A Comparative Analysis Put-Call Ratio Vs. Volatility Index. Journal of Business Economics Research (JBER), 6(8). https://doi.org/10.19030/jber.v6i8.2458

Published in August 2008 in the Journal of Business Economics Research, the journal article titled "Measure of Investor Sentiment: A Comparative Analysis Put-Call Ratio Vs. Volatility Index" by authors Jones and Bandopadhyaya explore the ability of two indicators' ability to capture investor sentiment. The authors motivation of the study is to explore investor sentiment as a possible determinant of asset prices and how "investor sentiment may explain short-term movements in asset prices better than any other set of fundamental factors" (Jones and Bandopadhyaya, 2008). The horizon of the study spans from January 2, 2004 through April 11, 2006 and uses data compiled from the Chicago Board Options Exchange.

The first indicator the authors explore is the put-call ratio of daily options volume on the SP 500, citing previous research that suggests as the put-call ratio of the SP 500 rises (falls), the market is likely to sell-off (rally). The second indicator the authors explore is the VIX, which is often considered to be "the world's premier barometer of investor sentiment and market sentiment" (Jones and Bandopadhyaya, 2008). The relationship between the VIX and the market is expected to be the same as the relationship between the put-call ratio and the market. A rising VIX suggests that there is a heightened sense of fear and uncertainty among market participants, and as a result the market will fall as investors look to unwind their positions and reduce their exposure.

The authors first use "a random walk-model to see what portion of the variability in the daily movement of the SP 500 index is explained by past values of the index itself" (Jones and Bandopadhyaya, 2008). Whistling the same tune as the efficient market hypothesis of Fama-French, the authors argue that past values of the SP 500 are expected to reflect all relevant and available economic information that could affect the index "especially

if the data are high frequency" (Jones and Bandopadhyaya, 2008). The unexplained portion (i.e. the residuals) from regressing the SP 500 on its previous day's value is then "a result of other non-economic related factors such as changes in market sentiment". The authors report regression results including the variables, coefficients, test statistics and p-values from regressing both the put-call ratio and the VIX on the residuals from the initial random walk model. The authors report high statistical significance for both the put-call ratio and the VIX, concluding that the put-call ratio contains' greater explanatory power and statistical significance in comparison to the VIX.

2.1.2 Measuring Investor Sentiment in Equity Markets

Bandopadhyaya, A., Jones, A. Measuring investor sentiment in equity markets. J Asset Manag 7, 208–215 (2006). https://doi.org/10.1057/palgrave.jam.2240214

In the early 2000's, authors contributing to the behavioral finance literature seem to imply and suggest that in some instances swings in investor sentiment appear to better explain fluctuations in asset prices than short term fundamentals such as price-to-earnings and book-to-equity ratios. In the paper titled "Measuring Investor Sentiment in Equity Markets" published in February 2006, authors Jones and Bandopadhyaya develop an equity stock market sentiment index to explore "how this measure can be used in a stock market setting by studying the price movements of a group of firms which represent a stock market index".

To construct an equity market sentiment, authors Bandopadhyaya and Jones build upon the previous work of Persaud (1996) to develop a measure of the general market's attitude or appetite towards risk. Using daily return data from July 2003 to July 2004, daily returns are calculated for all securities contained in the Massachusetts Bloomberg Index, which is a "price-weighted index designed to measure the performance of the Massachusetts Economy" (Bloomberg). Through use of the average standard deviation of the past 5 daily returns for each security contained in the MBI, "the daily rate of return and the historic volatility are ranked, and the Spearman rank correlation correlation coefficient between the rank of the daily returns for each firm and the rank of the historic volatility of the returns for each firm is computed, and the result is multiplied by 100". The authors find that fluctuations in the constructed equity market sentiment are significantly correlated with news flow concerning securities contained within the MBI.

After constructing the equity market sentiment index, the authors regress the one-period lagged returns of the MBI and the current period return in the equity market sentiment index against the current period return of the

MBI. An important finding of interest is that "while the lagged value of the return in MBI has an insignificant impact on the dependent variable MBI, the coefficient on the equity market sentiment index is highly significant". This finding suggests that daily returns in the MBI on any given day are primarily driven by investors' appetite for risk, not the previous day's returns in the MBI. The authors conclude the article very clearly by stating that "researchers and practitioners should pay close attention to investor sentiment as a determinant of changes in financial markets".

2.1.3 Investor Sentiment in the Stock Market

Baker, M Wurgler, J. (2007). "Investor Sentiment in the Stock Market". The Journal of Economic Perspectives , Spring, 2007, Vol. 21, No. 2 (Spring,2007), pp. 129-151. American Economic Association https://www.jstor.org/stable/30033

In the paper titled "Investor Sentiment in the Stock Market" published in the spring of 2007, authors Baker and Wurgler define investor sentiment as "a belief about future cash flows and investment risks that is not justified by the facts at hand" presently available to an investor. Prior to periods of prolonged positive investor sentiment in recent years such as the lead up of the dotcom bubble crash in the late 1990's, it was not evident that investor sentiment had a material impact on asset prices. The literature has shifted in recent years from asking whether investor sentiment has any material effect on asset prices to determining how to quantity and measure investor sentiment using a systematic, numerical approach.

The paper intends to explore which stocks are most likely to be affected by swings in investor sentiment. Authors Baker and Wurgler propose an initial hypothesis outlining that "stocks of low capitalization, younger, unprofitable, high-volatility, non-dividend paying, growth companies or stocks of firms in financial distress are likely to be disproportionately sensitive to broad waves of investor sentiment" as these stocks tend to be more difficult to value and as a result more difficult to arbitrage. As arbitrageurs are less likely to actively trade these stocks back to prices more representative of their fundamental values, their prices are more likely to wander off due to positive or negative investor sentiment. More simply put, the underlying idea is stocks that are speculative by nature and more difficult to value will have higher relative valuations during periods of heightened investor sentiment.

Prior to outlining their chosen proxies of investor sentiment, authors Baker and Wurgler write that "investor sentiment is not straightforward to measure, but there is no fundamental reason why one cannot find imperfect proxies that remain useful over time", suggesting that the more practical approach is to combine several imperfect measures into one aggregate investor

sentiment index. A variety of sensible investor sentiment proxies are reviewed including investor surveys, investor mood, retail investor trades, mutual fund flows, trading volume, dividend premium, closed-end fund discount, option implied volatility, first-day IPO returns, IPO volume, equity issues over total new issues and insider trading. Due to data availability, a sentiment index is constructed based on the same six proxies used in Baker and Wurgler (2006) (trading volume as measured by NYSE turnover; the dividend premium; the closed-end fund discount; the number and first-day returns in IPOs; and the equity share in new issues). As a result of the deregulation of brokerage commissions and persistent decline in trading costs leading to an upward trend in turnover, "the log of turnover minus a five-year moving average" is used. The authors are also careful to isolate any influence of economic fundamentals by regressing each proxy on a set of macroeconomic indicators. The residuals of the regressions are then used as the sentiment proxies.

Using monthly mutual fund flows data from The Investment Company Institute, principal component analysis is used to "detect general patterns across several time series while ironing out distracting idiosyncratic fluctuations". Mutual funds tend to reveal aggregate decisions of a large set of investors who are generally less sophisticated and more likely to exhibit sentiment driven investment behavior. The resulting correlation of 0.36 between the speculative demand contained in the second principal component of the mutual fund flows and the sentiment index is highly significant implying that Baker and Wurgler sentiment index based on the six proxies outlined earlier "to a large extent captures a prevailing "greed" versus "fear" or "bullish" versus "bearish" notion". Additional significant findings revealed in the paper are that when sentiment is high, subsequent market returns are low as well as that the impact of market sentiment is stronger for smaller stocks with more volatile monthly returns (i.e., more difficult to arbitrage).

2.1.4 Investor Sentiment Measures

Qiu, L. X., Welch, I. (2006). Investor Sentiment Measures. Social Science Research Network. https://doi.org/http://dx.doi.org/10.2139/ssrn.589641.

In the paper titled "Investor Sentiment Measures" published July 2006 in the Social Science Research Network, authors Qiu and Welch attempt to validate two widely used proxies to empirically quantify and measure general investor sentiment. The two measures are the closed end fund discount and consumer confidence, two indicators often used as proxies in the behavioral finance literature. Referring to previous literature by authors Lee, Shleifer and Thaler (1991), Qiu and Welch interpret the closed-end fund discount measure as having a negative correlation with investor sentiment. A closed-

end fund (hereafter referred to as CEF) is a type of investment fund that raises a fixed amount of capital through an initial public offering. Contrary to open-end funds that issue and redeem shares based on the net asset value of the underlying assets contained within the fund, CEFs issue a fixed number of shares that are then listed and traded on an exchange. As a result, the share price of a CEF is determined by traditional supply and demand dynamics allowing it to trade at a discount or premium to its net asset value. The CEF discount refers to the difference in percentage terms between the market price of a CEF share and its net asset value. The second proxy for investor sentiment is one of the components in the UBS/Gallup's Survey of Investor Sentiment, which authors Qiu and Welch "believe to be the best available empirical direct proxy for investor sentiment". Using CEF discount as a proxy for investor sentiment strongly relies on the widely accepted empirical observation that CEFs are primarily held by less informed retail investors.

The paper offers three significant findings. The first being that the CEF discount is not a good measure of investor sentiment and has no correlation with the results of the UBS/Gallup measure of investor sentiment. Secondly, the authors show that the changes in consumer confidence (Michigan Consumer Confidence Index provided by Michigan Consumer Research Center) "correlate strongly with changes in the UBS/Gallup proxy". The consumer confidence proxy and UBS/Gallup sample different respondents (UBS/Gallup samples investors with wealth in excess of $100,000 while consumer confidence samples investors with wealth under $100,000), suggesting that the shared common factor is investor sentiment. Lastly, the authors explore the relationship between investor sentiment and markets. Qiu and Welch provide evidence outlining that "there is no difference in how poor and wealthy investor sentiment changes month-to month", indicating that wealth is not a determinant of investor sentiment changes.

The underlying intuition behind the empirical analysis of the paper is that since small decile stocks "are disproportionately held by noise traders, sentiment changes should change the spread between small decline firms and large decile firms". Additionally, the authors propose that changes in investor sentiment "disproportionately influence stocks not held by institutional but retail investors, especially if these stocks have insufficient liquidity to allow arbitrageurs to impose rational pricing." The results show that changes in the consumer confidence and CEF discount series have strong explanatory power in spread between the smallest and largest capitalization stocks held within the frequently used CRSP portfolios.

Towards the end of the paper, the authors caution readers to interpret the evidence presented with extra care as there exists "considerable academic sentiment when it comes to interpretations of investor sentiment". Until the

academic literature is presented with justifiable theories outlining explicitly and quantitatively how much investor sentiment should influence prices in financial markets, "no paper can fully confirm or reject one of the two perspectives in favor of the other" (the two perspectives being the classical rational investor theory and the behavioral irrational investor theory perspective).

Relevant: Yet, absent precise quantitative theories of how sentiment should influence financial markets in an irrational-world perspective vs. a rational-world perspective— so that we can attribute correlation that can be claimed by either perspectives exclusively to one or the other—we cannot determine whether investor sentiment is an entirely behavioral, irrational or a classical, rational phenomenon.

2.1.5 Investor Sentiment and the Cross-Section of Stock Returns

Baker, M Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. The Journal of Finance, Vol. LXI, no. 4, August 2006.

Published in August of 2006, the paper titled "Investor Sentiment and the Cross-Section of Stock Returns" written by authors Baker and Wurgler intends to "study how investor sentiment affects the cross-section of stock returns". The idea behind the paper stems from the assumption that investor sentiment related shocks are more observable in securities that are difficult to value and as a result, difficult to arbitrage. A security whose price is difficult to value may be a newly listed firm with limited earnings history, a firm with an abstract product that is difficult to value, a non-dividend paying firm or a firm whose valuation tends to be highly subjective. Authors Baker and Wurgler initially provide a theoretical prediction which states that investor sentiment shocks have cross-sectional effects, or more simply put do not impact all security prices equally in a uniform fashion.

In order for authors Baker and Wurgler to empirically measure and quantify investor sentiment, a number of proxies are considered and used to create various time-series conditioning variables that are then aggregated into a composite sentiment index based on the variables first principal component (principal component analysis). To test the cross-section of stock returns and the composite stock index, using monthly stock returns in the period of 1963 to 2001, the authors form "equal-weighted decile portfolios based on several firm characteristics" such as market capitalization, age of firm, return volatility, profitability and dividend paying vs. non-dividend paying firms. The authors report that "when sentiment is low, subsequent returns are higher on very young (newly listed) stocks than older stocks, high-return volatility than low-return volatility stocks, unprofitable stocks than profitable ones, and non-payers than dividend payers", and that "when sentiment is high,

these patterns completely reverse". The intuition behind these findings is that during a bubble period where sentiment is irrationally high, the propensity to speculate on firms with highly subjective valuations is high, whereas an older firm with long earnings history is less subjective and as a result "less likely to be affected by fluctuations in the propensity to speculate".

Authors Baker and Wurlger offer up the definition of investor sentiment as "optimism or pessimism about stocks in general", which is a very simplistic, concise and all-encompassing definition to say the least. In the third section of the paper titled "Empirical Approach and Data", the methodology is outlined on the basis of understanding that investor sentiment frequently leads to patterns of mispricing in securities, specifically securities that are difficult to value and as a result, difficult to arbitrage. With regards to quantifying investor sentiment, the authors use six proxies while stating that "there are no definitive or uncontroversial measures". The six proxies include the "closed-end fund discount, NYSE share turnover, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium". Using principal component analysis, the authors isolate the sentiment component from the idiosyncratic/non-sentiment component from each of the six proxies.

CONTINUE ABOVE FROM PAGE 17

2.1.6 Firm-specific investor sentiment and daily stock returns

Sang Ik Seok, Hoon Cho, Doojin Ryu. 2019. Firm-specific investor sentiment and daily stock returns. The North American Journal of Economics and Finance, Volume 50. DOI: https://doi.org/10.1016/j.najef.2018.10.005

Published in the North American Journal of Economics and Finance in November of 2018, the journal article titled "Firm-specific investor sentiment and daily stock returns" by authors Seok, Cho and Ryu explores the relationship between investor sentiment and daily stock returns in the Korean stock market. As a result of "Korea's high degree of collectivism" (Seok et al), the influence of market sentiment is increased. Additionally, there is a strong level of participation among individual investors "who are usually uninformed, noisy, and sensitive to market sentiment" (Seok et al, 2019). Given these characteristics of the Korean stock market, a study pertaining to market sentiment has the potential to generate revealing results.

The authors note that much of the existing literature analyzing market sentiment makes use of relatively low frequency data (i.e. monthly or annually), and often conclude that periods of high investor sentiment are often followed by lower subsequent stock returns. Authors Seok et al cleverly point out that "low frequency measures cannot effectively capture the mispricing

process and cannot correctly estimate the time frame of the mispricing" and that through the use of analyzing short term frequency data (i.e. daily or weekly), one can determine whether mispricings due to swings in investor sentiment are immediately corrected or achieve an understanding of how long the mispricing lasts.

Firms are sorted on a variety of characteristics such as firm size, stock volatility, profitability and growth opportunities in order to determine if firms that are difficult to arbitrage and value are more sensitive to swings in investor sentiment. To quantify firm specific investor sentiment, authors Seok et al use a similar approach taken by authors Baker and Wurgler 2006. Using principal component analysis, the authors combine the relative strength index (RSI), psychological line index (PLI), adjusted turnover rate (ATR) and logarithm of trading volume (LTV) into a firm specific sentiment index. It's important to note that LTV and ATV are both constructed using trading volume, and begs the question if it is better to include one or the other instead of both. Additionally, both the RSI and PLI are constructed using price data, and as mentioned previously in the introduction of this paper, are indirect, second order measurements of investor sentiment. The authors intend to "identify instances when the trading volume is high for no rational reason" (Seok et al, 2019). Using a similar approach to Baker and Wurgler 2006, the raw proxies are individually orthogonalized to three factors (firm size, book to market ratio, earnings-price ratio) to control for the effect of a firm's fundamentals on returns. The authors note that "the residuals from these regressions are cleaner proxies for investor sentiment" (Seok et al, 2019).

In the empirical results section, the authors report that the sign of the coefficient on sentiment is significantly positive, suggesting that at the firm level in the Korean stock market, periods of high investor sentiment are subsequently followed by positive returns in the short term. The authors also report findings that indicate the influence of investor sentiment is significantly stronger for smaller firms. The explanation offered is that arbitrageurs in the Korean stock market are not able to effectively offset the mispricings occurring as a result of swings in investor sentiment in smaller, more volatile, less profitable firms. The authors conclude by reiterating the findings of Baker and Wurgler 2006, confirming that publicly traded Korean firms that are more difficult to arbitrage and as a result more difficult to value are more strongly affected by investor sentiment.

2.2 Investor Attention

2.2.1 In Search of Attention

Da, Z Engelberg, J Gao, P. (2011). "In Search of Attention". Wiley for the American Finance Association. https://www.jstor.org/stable/41305167

Published in the Journal of Finance in October of 2011, the article titled "In Search of Attention" written by authors Da, Engelberg and Gao appears to be the first paper that explores the relationship between stock prices and Google's search frequency data platform called Google Trends. Google Trends is a web-based tool offered by Google allowing users to observe the popularity of a specified search term throughout a specified period of time. Users are able to extract the relative search frequency of searches or phrases varied by regions, languages and time periods. Using the search frequency data, authors Da, Engelberg and Gao construct a search volume index (hereafter referred to as "SVI") to analyze the correlation between the proposed proxy of investor attention and Russell 3000 stock prices.

The paper begins by outlining that investor attention is a limited and scarce resource and attributes the initial idea of the paper stemming from recent studies that "provide a theoretical framework in which limited attention can affect asset pricing statics as well as dynamics". Traditional indirect proxies of investor attention such as headline news contained in a day's issue of the Wall Street Journal do not guarantee attention unless investors actually allocate their attention accordingly and read it. The authors of the paper define abnormal SVI as "the log of SVI during the current week minus the log median SVI during the previous eight weeks", and find that "majority of the time-series and cross-sectional variation in ASVI remains unexplained by alternative measures of attention" and that a given stock's SVI "has little correlation with a news-based measure of investor sentiment". Additionally, by studying the changes in equity turnover between trading venues that typically attract less sophisticated investors and venues that attract more sophisticated investors, the authors find and suggest that "SVI likely captures the attention of less sophisticated investors". The paper also reviews investor sentiment referencing prior work by authors Barber and Odean (2007), but the authors state "it is not clear how investor attention and sentiment should be related to each other".

Through the use of vector auto-regressions across four different constructed variables, authors Da, Engelberg and Gao are able to show that "SVI captures investor attention in a more timely fashion than extreme returns or news", suggesting that Google Trends is a better and more direct measure of investor attention. When regressing existing proxies of investor

attention such as turnover or news coverage, a resulting R-squared of approximately 3.3% suggests that the conventional investor attention proxies only explain a small portion of the variation in ASVI. The authors also use SVI to test the validity of the price pressure hypothesis of Barber and Odean (2008), and "find that an increase in SVI for Russell 3000 stock predicts higher stock prices in the next 2 weeks and an eventual price reversal within the next year".

Another interesting and relevant finding outlined in the paper titled "In Search of Attention" is the negative coefficient between two variables constructed by authors Da, Engelberg and Gao. The two variables are "log market cap" and "ASVI", which intend to measure the relationship between the magnitude of changes in ASVI relative to the market cap of a firm included in the Russell 3000. The negative coefficient between these two variables suggests "a larger price increase following an increase in ASVI among smaller Russell 3000 stocks". Google Trends data generally captures the attention of less sophisticated retail investors as more sophisticated investors have access to better platforms such as a Bloomberg Terminal. In tandem with the evidence that in the years 2004 to 2007 (the sample period of the paper) retail investors tend to gravitate towards smaller stocks "the positive price pressure is only present among the smaller half of (their) Russell 3000 stock sample". It's interesting to note that between 2007 and the 1st quarter of 2023 (the time of writing), fractional share offerings on many retail investor platforms have allowed less sophisticated and wealthy investors to participate in trading larger stocks.

2.2.2 Market Liquidity As A Sentiment Indicator

Baker, M., Stein, J. C. (2004). Market liquidity as a sentiment indicator. Science Direct, 7(3), Pages 271-299. https://doi.org/10.1016/j.finmar.2003.11.005.

In the paper titled "Market liquidity as a sentiment indicator" published in 2004 by authors Baker and Stein, an alternative, theoretical contribution is made to attempt to explain and analyze the relationship between market liquidity and expected returns. The fundamental idea behind the literature is that "in a world with short-selling constraints, market liquidity can be a sentiment indicator". The paper attempts to understand "why time-variation in liquidity, either at the firm level or for the market as a whole, might forecast changes in returns". Two assumptions are explicitly made, one concerning market liquidity and the other about individual investor behavior. The former assumes that investors are subject to and limited by short selling constraints. The latter proposes the "existence of a class of irrationally overconfident investors" who overweight the value of their own private trading

signals. Due to the market friction assumption that refers to short selling constraints, individual irrational investors will only be active in the market when the valuations inferred from their private signals are higher than the valuation of rational investors, "i.e. when their sentiment is positive and when the market is, as a result, overvalued" and "when the sentiment of irrational investors is negative, the short-selling constraints keeps them out of the market altogether". The authors suggest that measures of market liquidity (such as trading volume, market turnover rate) can provide an "indicator of the relative presence or absence of these investors, and hence the level of prices relative to fundamentals".

The positive relationship between market liquidity and investor sentiment is outlined and conveyed through visual representations of empirical results performed on US value-weighted and equally weighted equity portfolios contained in the CRSP database. Appendix X contains the development of price action, participation of 'smart' and 'retail' investors' and market liquidity. In region 1, 'smart investors dominate the market' and 'w' (a variable representing the price impact of an individual trade) is increasing and 'hence liquidity and trading volume are low'. In region 2, both retail and smart investors are participating in the market and added additional provision of liquidity from retail investors reduces price impact. In region 3, retail investors 'dominate the market' with 'smart traders on the sidelines, w is low and hence liquidity and trading volume are high'. To summarize, as retail investors increasingly participate in region 2 and 3 as a result of becoming more optimistic about the potential outcome of a trade, 'not only do liquidity and trading volume increase, but expected returns fall'. Despite retail investors subject to the short-selling constraints assumption outlined earlier in the paper, the authors indicate that as retail investors become increasingly optimistic and are doing all the buying, 'smart investors continue to exert the same marginal influence on price by taking short positions'.

Relevant: Many of the effects in our model are vividly illustrated by the behavior of Internet stocks during the boom period from January 1998 to February 2000. Ofek and Richardson (2003) document that the extraordinarily high valuations in this sector at this time were accompanied by very low bid–ask spreads and unusually high trading volume.

2.2.3 The Effect of Social Media on Trading Behavior: Evidence From Twitter

3 Theoretical Framework

3.1 Individual & Institutional investors

In order to understand the market, it is important to understand the different actors on the market, in this case the investors that trade and invest in securities. You tend to divide the investors in two groups, the individual or private investors and the institutional investors. While both groups are a kind of investor they differ in their approach to trading and their ability to trade and analyze data. While individual investors could be anyone who trades through a broker or stock exchange, institutional investors belong to large institutions like funds or banks that trade on the behalf of their beneficiaries. Due to this institutional investors tend to commit to more analysis and research before acting on a trade. Institutional investors tend to hold more knowledge about the financial markets and companies in general making them less susceptible to missing key data. They also tend to spend more time on researching companies compared to the individual investors as they are employed by the institutions to make money for their clients. Due to this individual investors tend to be less informed and therefore more likely to act on news or social media posts.(Durbin)[4]

Individual investors are hinted at to have a tendency of being overconfident in their ability to predict the market as well as having access to asymmetric information. There is also the risk of sensation seeking that individual investors not constrained to trade with the best interests of their clients in mind, tend to pursue. Due to all this individual investors tend to be less informed and therefore more likely to act on news or social media regarding posts, chasing the action trying to make a profit. (Barber) [5]

3.2 Principal Component Analysis

Principal Component Analysis (hereafter referred to as PCA) is a statistical technique used frequently in the formation of firm-specific investor sentiment and investor attention indices. PCA can be used to reduce the dimensionality of a multivariate dataset using the following five steps:

1. Standardizing the raw proxies to ensure unit variance across all variables.

2. Calculating the covariance matrix of the standardized proxies.
3. Calculating the eigenvalues and eigenvectors of the covariance matrix.
4. Creating a feature vector to decide which principal components to keep.
5. Recasting the original data over the feature vector. (Reference: Jaadi Zakaria, pca)

3.2.1 Standardizing

Standardization of data is very useful when dealing with data where the variables differ a lot in what range their values lie in, either due to being measured in different units or because the values take different characteristics. This is important as to not let certain variables dominate the data due to them being measured differently. When standardizing data, the z-score is calculated for each point, for each variable in the dataset. The z-score is calculated by subtracting the mean of the variable from each point and then dividing by the variable's standard deviation.

$$x_{standardized} = \frac{x_i - \mu_{variable}}{\sigma_{variable}} \tag{3.1}$$

This is done to bring down all variables to the same scale as all variables get a mean of 0 and a standard deviation of 1. Once this is completed the variables in the dataset can be compared to each other in a fair way that won't favor any variables more than the others. Without this crucial step it would not be possible to complete the principal component analysis. (Jaadi Zakaria, standardization)

3.2.2 Covariance Matrix

Once the data has been standardized it is possible to compute a covariance matrix of all the covariances between the different variables. The covariance between two variables of equal size is measuread as;

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \tag{3.2}$$

Once you have calculated all the covariances between all the variables you can compute a matrix of m **x** n dimensions where $m = n$ and where n is the amount of variables you include in the covariance matrix.

$$C_{nXn} = \begin{pmatrix} COV(X_1, X_1) & COV(X_1, X_2) & \cdots & COV(X_1, X_n) \\ COV(X_2, X_1) & COV(X_2, X_2) & \cdots & COV(X_1, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ COV(X_n, X_1) & COV(X_n, X_2) & \cdots & COV(X_n, X_n) \end{pmatrix}$$

Following step 2 and computing a covariance matrix from n amount of variables will result in a square covariance matrix with the dimensions n **x** n. From this covariance matrix it is possible to move on to step 3 of the PCA and compute the covariance matrix's eigenvalues and eigenvectors (Linda Smith).

3.2.3 Eigenvalues & Eigenvectors

An eigenvalue to a vector is denoted as λ_i and is a value for which the following is true:

$$AX = (E \cdot \lambda)X \rightarrow (A - E\lambda)X = 0 \rightarrow A - E\lambda = 0 \tag{3.3}$$

where E is the unity matrix of n **x** n dimension for which $EA = AE = A$ is true. The unity matrix is displayed with only ones in the diagonal:

$$E = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \rightarrow E \cdot \lambda = \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{pmatrix} = \lambda$$

This means that 3.3 can be displayed as

$$\begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{1,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{1,1} & \cdots & A_{n,n} \end{pmatrix} - \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{pmatrix} = \begin{pmatrix} A_{1,1} - \lambda & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} - \lambda & \cdots & A_{1,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{1,1} & \cdots & A_{n,n} - \lambda \end{pmatrix}$$

The eigenvalues can be found by calculating the determinant of 3.3 and setting it equal to 0. Once the eigenvalues have been found you can calculate the corresponding eigenvectors for each eigenvalue by solving the system of equations created by each eigenvalue when inserted into 3.3.(Kg Andersson, lienär algrebra)

Once the covariance matrix has been created in step 2 it is possible to find its eigenvalues. There will be the same amount of eigenvalues as amount of variables that the covariance matrix for has been calculated the. With these values it will be possible to find their corresponding eigenvectors and move on to step 4.

3.2.4 Choosing our Feature Vector

Once the eigenvalues and eigenvectors have been computed it is time to choose what principal components (PCs) to include and which to exclude. The greater a PC's eigenvalue, the greater proportion of variance explained. The variance of a PC is equivalent to the respective eigenvalue divided by the sum of all eigenvalues.

$$VAR_{PC_i} = \frac{\lambda_i}{\sum \lambda_i} \tag{3.4}$$

Since the eigenvalues have been computed from a covariance matrix of a standardized data set, the sum of all eigenvalues is equal to the amount of variables as well as eigenvalues, $\sum_{i=1}^{n} \lambda_i = n_\lambda = n_{variables}$.

The first principal components are the values within the eigenvector corresponding to the first eigenvalue, the second principal components are the values within the eigenvector corresponding to the second eigenvalue, so on and so forth. The principal components explaining the most variance are picked, meaning the principal components belonging to the highest eigenvalues are picked. In many cases only the principal component belonging to the maximum eigenvalue is picked as the final principal component and become the feature vector.

3.2.5 Final Step

Once the feature vector has been decided, the final step to complete the the PCA is to multiply the original standardized data with the feature vector and its loadings. With the data now transformed the PCA is completed .

3.3 Measuring Social Attention

3.4 Measuring Social Sentiment

3.5 Measuring Investor Attention

3.6 Measuring Investor Sentiment

3.7 Stationarity

In time-series analysis, stationarity must be taken into consideration. In this study, we regress our explanatory (independent) variables including the calculated sentiment and attention indices on our dependent variables including daily, weekly and monthly stock returns. In the simplest of terms, stationarity is a measurement of how much a time-series wanders off from its current path. A time-series is considered stationary if it has a constant mean and variance over time. On the other hand, a non-stationary time-series contains a unit root indicating that current observations are relatively dependent on past observations. Feeding non-stationary time-series data into a regression model can lead to spurious results and overstated statistical significance. (Lena Jaroszcek, Empiral Finance PDFs, Chapter 3). The stationarity of a time-series can be confirmed by checking the three followings conditions:

1) $E(y_t) = \mu \rightarrow mean\, is\, finite\, and\, constant\, across\, t$
2) $Var(y_t) = \sigma^2 \rightarrow variance\, is\, finite\, and\, constant\, across\, t$
3) $Cov(y_t,\ y_{t-1})\ for\ s \neq 0\, is\, finite\, and\, a\, function\, of\, s\, but\, not\, of\, t$
$(Lena Jaroszcek, Empirical Finance PDFs, pg 9 chapter 3)$

Checking for stationarity is critical in time-series analysis as our models implicitly assume that current values are independent of previous values. If our data is non-stationary suggesting that previous values have an effect on current on values (i.e., contains a unit root), regression models may lead to spurious results. This study intends to produce robust and reliable empirical results and therefore the dangers of non-stationarity need to be taken into consideration, and dealt with accordingly where necessary. There exists multiple ways to overcome the dangers of non-stationarity in time-series analysis such as logarithmic transformation or differencing. Logarithmic transformation consists of taking the natural logarithm of all values in a time-series. Differencing transforms the data by taking the difference between the current and previous value, $\Delta x(t) = x_t - x_{t-1}$.
When transforming data using differencing, any inherent trend or seasonality

is removed from the time-series, resulting in a more stable mean over time. If transformation via differencing does not suffice in rendering a time-series stationary, further transformation using logarithmic transformation may be required. It is important to highlight that logarithmic transformation must always be followed by differencing.(REFERENCE: Rayhaan Rasheed)

To check for stationarity in a time-series, two tests are frequently used in the statistical literature. The first test is known as the Augmented Dickey-Fuller (ADF) test, and the second is the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. The ADF test is a parametric test which seeks to validate the existence of a unit root to confirm whether or not a time-series is stationary. The ADF tests also takes into account the presence of trend and auto-correlation in a time-series. The intuition behind the ADF test gives rise to the null hypothesis (H_0) that the time-series has a unit root and is non-stationary and an alternative hypothesis ((H_A) that the time-series does not have a unit root and is stationary. The KPSS test is a non-parametric test that specifically determines whether a time-series has a constant mean and variance over time. The null and alternative hypothesis of the KPSS test are the inverse of the ADF test, where the null hypothesis tests if a time-series is stationary and does not have a unit root while the alternative hypothesis tests if a time-series is non-stationary and has a unit root. (REFERENCE: Statsmodels).

The optimal method of verifying the existence of stationarity in time-series analysis is to use both the ADF and KPSS tests side by side in tandem as they are complementary to one another. As described previously, the ADF test considers the presence of trend and auto-correlation while the KPSS tests confirms if a time-series has a constant mean and variance over time. If each tests confirms that a time-series is non-stationary (stationary), it can be concluded that the time-series is in fact non-stationary (stationary). If the KPSS tests indicates stationarity while the ADF test does not, the time-series is said to be trend-stationary and some form of detrending transformation may be required. However, if the KPSS test indicates non-stationarity while the ADF test indicates stationarity, a time-series is said to be difference stationary and a differencing transformation may be be required. (REFERENCE: Statsmodels)

3.8 Orthogonal Data

3.8.1 Macroeconomic Data

Following the construction of the sentiment and attention proxies, it is important to consider that the "principal component analysis cannot distin-

guish between a common sentiment component and a common business cycle component" (Baker and Wurgler 2006). Fluctuations in macroeconomic and/or fundamental data has an affect on investor sentiment and this relationship is covered extensively in the literature (NEED REFERENCES HERE). In order to isolate the common sentiment component from the principal components, we follow a similar approach to Baker and Wurgler (2006) where each raw proxy used in the construction of the sentiment and attention indices is regressed on various monthly macroeconomic data. The monthly macroeconomic data includes growth in industrial production index (Federal Reserve Statistical Release), growth in consumer durables (FRED), non-durables (FRED) and services (FRED) as well as a dummy variable for NBER recessions (FRED). The residuals from these regressions may then be a better proxy for firm-specific investor sentiment not explained by fluctuations in the overall economy. The following model is used to predict the raw proxy value:

$$\begin{aligned} y_{pred} = \beta_1 \cdot IPNCONGD + \beta_2 \cdot PCES \\ + \beta_3 \cdot IPDCONGD + \beta_4 \cdot INDPRODI \end{aligned} \tag{3.5}$$

Using the above regression model as a predictor of the raw proxy value, the residual is obtained by subtracting the predicted value from the actual value of the raw proxy.

$$y_{res} = y_{actual} - y_{pred} \tag{3.6}$$

As a result of the macroeconomic data only available at a monthly interval, this process is repeated for each variable for each of the six stocks only on the data sets containing monthly data.

3.8.2 Random-Walk Model

Another method of determining what degree of the variation of a stock price is not explained by fundamental or macroeconomic factors is using a random-walk model. In this section, we follow a similar approach to Jones and Bandopadhyaya (2008). A significant degree of variation in current stock prices can be explained by previous stock prices, hence stock prices being non-stationary. This is consistent with the efficient market hypothesis (Fama, 1970) which suggests that markets are informationally efficient, instantaneously incorporating all readily available information into a stock's current market price. Authors Jones and Bandopadhyaya point out that "past values of the (S&P500) index itself capture all relevant economic information that

affects the contemporaneous index values" and "any unexplained portion of the daily movement in the (S&P500) index must then result from changes in other non-economic factors" (Jones and Bandopadhyaya, 2008). These residual fluctuations in price may be a result of investor sentiment and/or investor attention. In order to determine portion of a stock price is not explained by past values, we use the following linear regression model:

$$P_t = \beta_0 + \beta_1 P_{t-1} + Residual_t \tag{3.7}$$

Where P_t is the current price, P_{t-1} is previous price and the portion of variation not explained by the previous price is $Residual_t$. We can now regress $Residual_t$ on a proxy to determine what degree of the variation in price not explained by previous values is explained by $Proxy_t$ (to determine explanatory power) or $Proxy_{t-1}$ (to determine predictive power). This leads to the following linear regression models:

$$Residual_t = \beta_0 + \beta_1 Proxy_t + \epsilon_t \tag{3.8}$$

$$Residual_t = \beta_0 + \beta_1 Proxy_{t-1} + \epsilon_t \tag{3.9}$$

Where $Proxy_t$ or $Proxy_{t-1}$ is a proxy for investor sentiment or investor attention that may explain $Residual_t$.

4 Methodology

4.1 Empirical Approach

TO BE WRITTEN

4.2 Data Collection

In order to carry out this study, a variety of data is collected from multiple financial databases at daily, weekly, and monthly time frames. Three separate data sets including all of the raw sentiment proxies are created for each of the six stocks, resulting in a total of 18 stock and frequency specific datasets. All variables and database sources are outlined in the following section.

4.2.1 Price, Trading Volume, Shares Outstanding

Daily, weekly and monthly stock prices (P_t) are retrieved from the Bloomberg database. Stock prices are used to calculate our dependent variables including simple returns, $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$, logarithmic returns, $logR_t = ln(\frac{P_t}{P_{t-1}})$ and

the random-walk model residuals as outlined in section 3.8.2. Daily shares outstanding ($SHOUT_t$) and daily trading volume (TV_t) are retrieved from the North American Compustat database accessed through the Wharton Research Data Services platform. By summing the trading volume of each respective trading day in a specified week or month, weekly and monthly trading volume is manually calculated.

4.2.2 Market Turnover Rate

Market turnover rate is calculated by $MTR_t = \frac{TV_t}{SHOUT_t}$, representing trading volume in a specified period as a portion of total shares outstanding. Authors Baker and Stein (2004) suggest that market liquidity can serve as a sentiment indicator, arguing that "in the presence of short-sales constraints, high liquidity is a symptom of the fact that the market is dominated by these irrational investors, and hence is overvalued" (Baker and Stein 2004). Authors Yang and Zhou (2016) calculate adjusted turnover rate as $ATR_t = \frac{R_t}{|R_t|} \cdot \frac{TV_t}{SHOUT_t}$, suggesting that "the turnover rate cannot judge whether the investor sentiment is optimistic or pessimistic" (Yang and Zhou, 2016). Raw trading volume data is expected to be non-stationary, while market turnover rate is expected to be stationary possessing a relatively constant mean and variance over time. The market turnover rate at the weekly (monthly) time interval differs from daily market turnover rate as the aggregate trading volume in a given week (month) is used instead.

4.2.3 Volatility Midpoint

The daily volatility midpoint (VMP_t) is calculated as the daily midpoint of the implied volatility values on both put and call options with 30 days to expiration. Implied volatility values corresponding to 30-day options are used in order to mirror the utility of the VIX as a market wide sentiment indicator, albeit at a firm-specific level. The implied volatility data for is retrieved from the Ivy-DB US Options database accessed through the Wharton Research Data Services platform. The daily volatility midpoint is calculated using the below equation. This process is repeated for each day in this study's horizon to generate a firm-specific volatility midpoint time-series for all stocks in our sample. In order to match the volatility midpoint with our weekly and monthly data sets, the end of week or end of month volatility midpoint is used. For months ending on a Saturday or Sunday, the closest trading day volatility midpoint is used.

$$VMP_t = \frac{\sigma_{t,\,Put,\,30\,days} + \sigma_{t,\,Call,\,30\,days}}{2} \tag{4.1}$$

4.2.4 Put-Call Ratio

The put-call ratio (PCR_t) is calculated as the total put option volume divided by the total call option volume on stock i in trading period t (i.e. day, week or month). The options trading volume is retrieved from the Ivy-DB US Option Metrics database accessed through the Wharton Research Data Services platform. Similarly to the volatility midpoint, we repeat this process for all stocks in our sample to create frequency-specific and firm-specific put-call ratio time-series. Daily put-call ratios are calculated using the below formula.

$$PCR_t = \frac{V_{put\ options,\,t}}{V_{call\ options,\,t}} \tag{4.2}$$

When calculating the put-call ratio on a weekly (monthly) time interval, total option volume for each trading day in a specified week (month) is used. The weekly (monthly) put-call ratios are matched to the stock-specific weekly (monthly) data sets. Weekly and monthly put-call ratios are calculated using the below formula.

$$\overline{PCR} = \frac{\sum_{t=1}^{n} V_{put\ options,\,t}}{\sum_{t=1}^{n} V_{call\ options,\,t}} \tag{4.3}$$

4.2.5 Search Volume Index

Search volume index (SVI_t) is another manually constructed variable included in this study. Daily search volume data is retrieved from Google Trends for all stocks in our sample. Google Trends does not provide raw search frequency data in search units, but instead represents search frequency values indexed on the maximum value within a specified time frame. The maximum value (SVI_{max}) in a specified time frame is set to 100 while the remaining values take on a value of any integer between 0 and 100 as a function of their relative size to SVI_{max}.

The literature suggests using a stock's ticker as a search term as it is unambiguous and a unique identifier (Da et Al, 2011). However, SVI_t intends to capture the aggregate investor attention, including both experienced and informed investors as well as inexperienced and uninformed investors. Although an experienced investor will likely search for a stock ticker directly, an inexperienced or new investor may be more likely to search for a firm's name followed by "stock". To capture and aggregate informed and uninformed investor attention, both a stock's ticker and the phrase "company name" + "stock" are entered as search terms. As an example, for Amazon we look at the search frequency data for "AMZN" and "Amazon Stock". To

aggregate the two search terms into a single SVI value, the search frequency time-series for both search terms needs to be indexed to the same SVI_{max}, therefore we search for both terms simultaneously by using two search terms on the Google Trends platform.

In order to retrieve daily data from Google Trends corresponding to this study's horizon, data must be retrieved and aggregated from numerous smaller time frames. As the specified time period entered into the Google Trends platform increases, the data frequency decreases from daily, to weekly, to monthly. To ensure the multiple smaller time frames can be combined, we ensure they overlap by exactly one day. The last day included in the previous time frame is denoted LDP, while the first day included in the current time frame is denoted FDC.

$$t_{max,\,LDP} = t_{min,\,FDC} \tag{4.4}$$

The ratio between the search frequency value of the last day of the previous time frame and the first day of the current time frame is then given by:

$$Comparative\ Ratio = \frac{SVI_{t_{max,\,FDC}}}{SVI_{t_{min,\,LDP}}} \tag{4.5}$$

We multiply all search frequency values in the current time frame by the respective $ComparativeRatio$. This is to ensure that all search frequency values in the current time frame are indexed according to the index of the previous time frame. Since the raw, stock-specific SVI datasets contain two time-series each (i.e. a column for stock ticker search term and a column for "company name" + "stock" search term), a comparative ratio is calculated for each column. This process of calculating a comparative ratio and multiplying the search frequency values contained in the subsequent time frame is repeated for all time frames. This process generates a resulting column for each date with all values indexed to the same SVI_{max}. In order to generate weekly and monthly search frequency data values, we sum the search frequency values for the respective period and get a total SVI for each week or month. The script used to carry out process of creating the SVI time-series can be found in the python file titled "GoogleTrends"[2]

Due to the repeated indexing throughout the process of converting the raw search frequency data into the SVI time-series, values can occasionally become very small that they either show up empty or represented as < 1". These values are manually set to 0 except if they belong to $t_{max,\,LDP}$ or

$t_{min,\,FDC}$. If the value belongs to $t_{max,\,LDP}$ or $t_{min,\,FDC}$, they are manually set to < 0.01 and reflect the comparative ratio of the second column.

TO FINISH EDITING WITH VINCENT (END)

Due to the SVI only taking the value of an integer it can happen that it is so small that it either shows up as empty or just < 1". These values have been set to 0 with the exception that if they belong to either . In that case they been set to low values of < 0.01 and reflecting the ratio of the second column. Due to the fact that values can only be an integer it can also happen that the ratios for the two different columns do not equal each other. We opted to use different ratios for each column instead of an average of the two ratios for two main reasons. For most ratios there was none or very small difference between them and by using the ratio belonging to their respective column we ...

We acknowledge that there are some issues with the SVI from Google Trends, mainly that the values are indexed and can only take the value of an integer. Therefore we have been forced to remodel our data to overcome these issues as explained above. Despite the limitations of Google Trends we don't expect these issues or our workarounds to play any significant role in the overall purpose of the SVI in our studies.

TO FINISH EDITING WITH VINCENT (END)

4.2.6 Bloomberg Social Velocity Factors

The remaining proxies for investor sentiment and investor attention are retrieved from the Bloomberg database. As mentioned in the previous literature section of this paper, the Bloomberg Social Velocity (BSV) proxies include Twitter Positive Count (TPC), Twitter Negative Count (TNC), Twitter Publication Count (TC), News Positive Count (NPC), News Negative Count (NNC) and News Publication Count (NC).

4.2.7 Stationarity of Our Proxies

To check the viability of our variables we decide to compute stationarity tests for our raw variables. We will apply both the augmented dickey fuller test(ADF) as well the the Kwiatkowski-Phillips-Schmidt-Shin(KPSS) test so that we can crossvalidate the responses from both tests. We will test their stationarity at a 5% significance level and then determine if our variables are stationary or not for all 6 companies and for daily, weekly and monthly data.

When applying both the ADF and KPSS test to our raw variables in the daily interval some interesting results can be soon. For $Price_{Daily}$ both

tests for all companies show non-stationarity. For all BSV_{Daily} variables except NC_{Daily} all test for stationarity in the ADF test but not in the KPSS test. For NC_{Daily}, GOOGL tests positively for stationarity in the KPSS test but due to it being the sole company to do so we can state that generally NC_{Daily} tests negatively for stationarity in the KPSS test. The combined results hints towards these six variables being difference stationary. The tests for PCR_{Daily} is the same as for most of the BSV_{Daily} variables, implicating that it also is difference stationary. The stationarity test show similar results for VMP_{Daily} but for two companies, AMZN and NFLX the KPSS test proves stationarity. Since a majority of the companies test negatively for staionarity with the KPSS test we can assume that the VMP_{Daily} is difference stationary. The results for TV_{Daily} are mixed but a majority of the ADF tests prove stationarity and a majority of the KPSS tests prove non-stationarity, implicating TV_{Daily} being difference staionary. For SVI_{Daily} we see that no companies test positively for stationarity in the KPSS test while only half test positively for staionarity in the ADF test. We can therefore draw no true conclusion but assume that SVI_{Daily} is either non stationary or difference staionary. The ADF tests on MTR_{Daily} prove stationarity while only half of the KPSS prove stationarity meaning that MTR_{Daily} is either difference staionary

	AMZN	**AAPL**	**META**	**GOOGL**	**MSFT**	**NFLX**
Price - ADF	No	No	No	No	No	No
Price - KPSS	No	No	No	No	No	No
TPC - ADF	Yes	Yes	Yes	Yes	Yes	Yes
TPC - KPSS	No	No	No	No	No	No
TNC - ADF	Yes	Yes	Yes	Yes	Yes	Yes
TNC - KPSS	No	No	No	No	No	No
NPC - ADF	Yes	Yes	Yes	Yes	Yes	Yes
NPC - KPSS	No	No	No	No	No	No
NNC - ADF	Yes	Yes	Yes	Yes	Yes	Yes
NNC - KPSS	No	Yes	No	No	No	Yes
PCR - ADF	Yes	Yes	Yes	Yes	Yes	Yes
PCR - KPSS	No	No	No	No	No	No
VMP - ADF	Yes	Yes	Yes	Yes	Yes	Yes
VMP - KPSS	Yes	No	No	No	No	Yes
TV - ADF	Yes	No	Yes	Yes	Yes	Yes
TV - KPSS	Yes	No	No	No	Yes	No
TC - ADF	No	Yes	Yes	Yes	Yes	Yes
TC - KPSS	No	No	No	No	No	No
NC - ADF	Yes	Yes	Yes	Yes	Yes	Yes
NC - KPSS	No	No	No	Yes	No	No
SVI - ADF	No	Yes	Yes	No	No	Yes
SVI - KPSS	No	No	No	No	No	No
MTR - ADF	Yes	Yes	Yes	Yes	Yes	Yes
MTR - KPSS	Yes	Yes	No	No	Yes	No

Table 1: Testing daily data for stationarity with ADF and KPSS tests.

When applying both the ADF and KPSS test to our raw variables in the weekly interval we can see some different results compared to the daily interval. As with the daily interval, both tests for all companies for $Price_{Weekly}$ show non-stationarity. We see that for TC_{Weekly}, NC_{Weekly} and TPC_{Weekly} the majority of both the ADF and KPSS tests prove non-stationarity but for NPC_{Weekly} and TNC_{Weekly} a majority of ADF tests show stationarity indicating both variables to be difference stationary. For NNC_{Weekly} we see that only half of the KPSS tests on NNC_{Weekly} prove stationarity so we can only assume that NNC_{Weekly} is either difference staionary or stationary. The tests on weekly VMP_{Weekly} and PCR_{Weekly} is almost similar to the tests on daily data, with the exception of one company per variable for the ADF test, and we can conlude that both variables remain difference stationary. As with

the tests on daily data the tests for TV_{Weekly} are mixed but a majority of the ADF tests prove stationarity and a majority of the KPSS tests prove non-stationarity, showing that TV_{Weekly} remains difference stationary. For SVI_{Weekly} we see that no companies test positively for stationarity in the KPSS test and all but one test negatively for staionarity in the ADF test. We can therefore assume that SVI_{Weekly} is non stationary. MTR_{Weekly} perfor,s very similar as in the daily interval, showing that MTR_{Weekly} remains difference staionary.

	AMZN	**AAPL**	**META**	**GOOGL**	**MSFT**	**NFLX**
Price - ADF	No	No	No	No	No	No
Price - KPSS	No	No	No	No	No	No
TPC - ADF	No	No	No	No	Yes	No
TPC - KPSS	No	No	No	No	No	No
TNC - ADF	Yes	No	Yes	Yes	Yes	No
TNC - KPSS	No	No	No	No	No	No
NPC - ADF	No	Yes	No	Yes	No	Yes
NPC - KPSS	No	No	No	No	No	No
NNC - ADF	Yes	Yes	Yes	No	Yes	Yes
NNC - KPSS	No	Yes	Yes	No	No	Yes
PCR - ADF	No	No	Yes	Yes	Yes	Yes
PCR - KPSS	No	No	No	No	No	No
VMP - ADF	Yes	Yes	No	Yes	Yes	Yes
VMP - KPSS	Yes	No	No	No	No	Yes
TV - ADF	Yes	No	Yes	Yes	Yes	No
TV - KPSS	Yes	No	No	No	Yes	No
TC - ADF	No	No	Yes	No	No	No
TC - KPSS	No	No	No	No	No	No
NC - ADF	No	Yes	Yes	Yes	Yes	No
NC - KPSS	No	No	No	Yes	No	No
SVI - ADF	No	No	Yes	No	No	No
SVI - KPSS	No	No	No	No	No	No
MTR - ADF	Yes	Yes	Yes	Yes	Yes	No
MTR - KPSS	Yes	Yes	No	No	Yes	No

Table 2: Testing weekly data for stationarity with ADF and KPSS tests.

When applying both the ADF and KPSS test to our raw variables in the monthly interval we can see some different results compared to the daily

and weekly interval. As with both the daily and weekly interval, both tests for all companies for $Price_{Monthly}$ show non-stationarity. We see that for $TC_{Nonthly}$ and $TPC_{Monthly}$ the majority of both the ADF and KPSS tests prove non-stationarity, however for $NC_{Nonthly}$ half of the KPSS tests prove stationarity implicating that $NC_{Nonthly}$ is either stationary For $NPC_{Monthly}$ and $TNC_{Monthly}$ a majority of ADF tests show stationarity indicating both variables to be difference stationary. For $NNC_{Monthly}$ we see that only half of the KPSS tests on $NNC_{Monthly}$ prove stationarity so we can only assume that $NNC_{Monthly}$ is difference stationary. The majority of ADF tests on $PCR_{Monthly}$ prove stationarity while a majority of KPSS tests prove non-stationarity implicating that $PCR_{Monthly}$ is difference staionary. For $VMP_{monthly}$ we see that a majority of the ADF tests how stationarity while half of the KPSS tests show stationarity, indicating that $VMP_{monthly}$ is either stationary or difference stationary. As with the tests on daily data the tests for $TV_{monthly}$ are mixed but a majority of both the ADF and KPSS tests prove stationarity, showing that $TV_{monthly}$ is stationary. For $SVI_{monthly}$ we see that no companies test positively for stationarity in either the KPSS or ADF tests. We can therefore assume that $SVI_{monthly}$ is non stationary. MTR_{Montly} sees some change in the KPSS test results as a majority now proves stationarity and MTR_{Montly} becomes stationary in the monthly time interval.

	AMZN	**AAPL**	**META**	**GOOGL**	**MSFT**	**NFLX**
Price - ADF	No	No	No	No	No	No
Price - KPSS	No	No	No	No	No	No
TPC - ADF	No	No	No	No	Yes	No
TPC - KPSS	No	No	No	No	No	No
TNC - ADF	Yes	Yes	Yes	Yes	No	No
TNC - KPSS	Yes	No	Yes	No	No	No
NPC - ADF	No	Yes	Yes	No	Yes	Yes
NPC - KPSS	No	No	No	Yes	No	No
NNC - ADF	Yes	Yes	Yes	No	Yes	Yes
NNC - KPSS	No	Yes	Yes	No	No	Yes
PCR - ADF	No	No	Yes	Yes	Yes	Yes
PCR - KPSS	No	Yes	Yes	No	No	No
VMP - ADF	Yes	No	Yes	No	Yes	Yes
VMP - KPSS	Yes	No	No	Yes	No	Yes
TV - ADF	Yes	No	Yes	Yes	Yes	No
TV - KPSS	Yes	No	Yes	Yes	Yes	No
TC - ADF	No	No	No	No	No	No
TC - KPSS	No	No	No	No	No	No
NC - ADF	No	No	Yes	No	Yes	No
NC - KPSS	Yes	No	No	Yes	No	Yes
SVI - ADF	No	No	No	No	No	No
SVI - KPSS	No	No	No	No	No	No
MTR - ADF	Yes	Yes	Yes	Yes	Yes	No
MTR - KPSS	Yes	Yes	No	Yes	Yes	No

Table 3: Testing monthly data for stationarity with ADF and KPSS tests.

The results of the stationarity tests show that a majority of thge variables tend to be difference stationary or that they are either difference stationary or non stationary. This indicates that differencing might be necessary to perform on all variables before creating our attentiona and sentiment ratios.

4.3 Computing Our Sentiment & Attention Ratios

With all the variables collected or calculated it is time to compute the ratios on which this paper is about. The attention ratio(ATTN) will invlude the variables News Publication Count (NC), Twitter Publication Count(TC), Search Volume Index(SVI) and Market Turnover Rate(MTR). The final sentiment ratio(SENT) will include the variables News Positive Publication Count(NPC), News Negative Publication Count(NNC), Twitter Positive Publication Count(TPC), Twitter Negative Publication Count(TPC), Volatiliy Midpoint(VMP) and Put Call Ratio(PCR).

4.3.1 Detrending & Standardizing

First step towards attaining a final ratio is to detrend the raw data and remove some of the variables' non stationarity. The method used is differencing all variables by calculating the delta for each point, $\Delta x_t = x_t - x_{t-1}$ and replacing the value with the delta. Once the data has been differenced, the data can be standardized following the same process as in 3.1.

4.3.2 PCA on Attention & Sentiment

Once the data has been standardized it is ready for a principal component analysis where attention and sentiment is analyzed separately. Each dataset is duplicated in size as all lags of the dataset's variables are included as a column. These values are lagged one timeframe for each variable resulting in a new lagged dataset. For ATTN the lagged dataset would look something like;

DATE	TC	TC_{lag1}	NC	NC_{lag1}	SVI	SVI_{lag1}	MTR	MTR_{lag1}
t	TC_t	TC_{t-1}	NC_t	NC_{t-1}	SVI_t	SVI_{t-1}	MTR_t	MTR_{t-1}
t+1	TC_{t+1}	TC_t	NC_{t+1}	NC_t	SVI_{t+1}	SVI_t	MTR_{t+1}	MTR_t
t+2	TC_{t+2}	TC_{t+1}	NC_{t+2}	NC_{t+1}	SVI_{t+2}	SVI_{t+1}	MTR_{t+2}	MTR_{t+1}
t+3	TC_{t+3}	TC_{t+2}	NC_{t+3}	NC_{t+2}	SVI_{t+3}	SVI_{t+2}	MTR_{t+3}	MTR_{t+2}

Each ratio then goes through a PCA twice as the first step is to indicate if the ratio is to use a variable or its lag. The covariance matrix for the lagged dataset is computed with its eigenvalues and corresponding eigenvectors. The eigenvector with the highest eigenvalue is picked as the feature vector with loadings that the standardized lagged dataset is finally multiplied with. The results of the first PCA are two different functions to calculate the first ratio,

one for attention and one for sentiment. For attention the function from the first PCA, the first attention ratio(FAR) would be the following,

$$\begin{aligned} FAR_{t,Ti} = \beta_1 \cdot TC_t + \beta_2 \cdot TC_{t-1} + \beta_3 \cdot NC_t + \beta_4 \cdot NC_{t-1} \\ + \beta_5 SVI_t + \beta_6 \cdot SVI_{t-1} + \beta_7 \cdot MTR_t + \beta_8 \cdot MTR_{t-1} \end{aligned} \tag{4.6}$$

For sentiment the function from the first PCA, the first sentiment ratio(FSR) would be the following,

$$\begin{aligned} FSR_{t,Ti} = \beta_1 \cdot TPC_t + \beta_2 \cdot TPC_{t-1} + \beta_3 \cdot TNC_t + \beta_4 \cdot TNC_{t-1} \\ + \beta_5 NPC_t + \beta_6 \cdot NPC_{t-1} + \beta_7 \cdot NNC_t + \beta_8 \cdot NNC_{t-1} \\ + \beta_9 VMP_t + \beta_{10} \cdot VMP_{t-1} + \beta_{11} \cdot PCR_t + \beta_{12} \cdot PCR_{t-1} \end{aligned} \tag{4.7}$$

The computed ratios across the entire timeframe are created through matrixmultiplication between the dataset and the transpose of the principal component loadings.

$$Ratio_{first,t_0 \to t_n} = Dataset_{standardized\,\&\,lagged} \times PCLoadings_1^T \tag{4.8}$$

With the first ratio computed the next step is to decide which of the pair of variables to keep, the variable or its lag. The decision is made by comparing the absolute value of the correlation between the first ratio and the variable to the absolute value of the correlation between the first ratio and the variable's lag.

$$Max_{Corr} = Max(|Corr(Ratio_{first},\, X)|,\, |Corr(Ratio_{first},\, X_{lag1})|) \tag{4.9}$$

The one of the pair with the maximum correlation to the computed ratio is picked and the other one is dropped. Repeating this process for all variables and their lags that the ratio consists of results in half of variables and lags being dropped along with their data, creating an edited dataset. The PCA is repeated for the new dataset computing a new covariance matrix along with new eigenvalues and eigenvectors to the covariance matrix. The eigenvalues belonging to the highest eigenvalue are once again picked as the principal component loadings to be multiplied with the edited dataset creating a second ratio.

The last step is to optimize the second ratio by changing the loadings and creating a final ratio by maximizing the correlation between the final ratio and the second ratio with the constraint that the final ratio needs to have a standard deviation of 1. Once the loadings have been optimized the final ratio can be computed.

4.3.3 Final Attention Ratio

With the data has gone through two principal component analyses and one optimization phase the resulting ratios can be computed. Attention includes four separate variables and therefore its lagged dataset had 8 variables in total. The final attention ratio will be a combination of four of these eight variables, one from each pair of a variable and its lag. A final attention ratio could look something like,

$$ATTN_{t,Ti} = \beta_1 \cdot TC_t + \beta_2 \cdot NC_t + \beta_3 \cdot SVI_{t-1} + \beta_4 \cdot MTR_{t-1} \tag{4.10}$$

4.3.4 Final Sentiment Ratio

Sentimenmt includes six separate variables and therefore its lagged dataset had twelve variables in total. The final attention ratio will be a combination of six of these twelve variables, one from each pair of a variable and its lag. A final sentiment ratio could look something like,

$$\begin{aligned} SENT_{t,Ti} = \beta_1 \cdot TPC_t + \beta_2 \cdot TNC_{t-1} + \beta_3 \cdot NPC_t \\ + \beta_4 \cdot NNC_t + \beta_5 \cdot PCR_t + \beta_6 \cdot VMP_{t-1} \end{aligned} \tag{4.11}$$

4.4 Regression Models

In the analysis there will be a few possible regression models that will be used to explore if there is a relationship between changes in price price or returns and attention or sentiment. These models will be to used to analyze to what extent this may be true and if so how much of the change is due to attention or sentiment.

4.4.1 Linear Regression Models

The first part of the analysis will look at linear regression models to find a possible linear connection between either price or return and sentiment or attention. Before we will regress our variables we will have to explore the stationarity of each one to identify which variables are not suited for regressions and which are.

The first linear model is the change of price at time t, $\Delta P_t = P_t - P_{t-1}$ as a function of the attention and sentiment at time $t-1$. As price tends to be non-stationary we expect that our price variable will be non-stationary, making this model unsuitable for regressions.

$$\Delta P_t = \alpha + \beta_a \cdot ATTN_{t-1} + \beta_s \cdot SENT_{t-1} \tag{4.12}$$

The second linear model will explore the linear connection between return at time t and attention and sentiment at time $t-1$. If there were any issues with the stationarity with the price variable, calculating the returns should have fixed that issue. There is some worry regarding the stationarity of the attention and sentiment ratios which might make this model unsuitable for regressions.

$$R_t = \alpha + \beta_a \cdot ATTN_{t-1} + \beta_s \cdot SENT_{t-1} \tag{4.13}$$

The third linear model is similar to 4.13 but instead of looking at sentiment and attention at time $t-1$ it considers the change in sentiment and attention at $t-1$, $\Delta Ratio_{t-1} = Ratio_{t-1} - Ratio_{t-2}$. By differencing both attention and sentiment by calculating the change, any possible issues with difference stationarity.

$$R_t = \alpha + \beta_a \cdot \Delta ATTN_{t-1} + \beta_s \cdot \Delta SENT_{t-1} \tag{4.14}$$

The fourth linear model builds on 4.14 but instead of taking the difference, it considers the percentage difference, $\%\Delta Ratio_{t-1} = \frac{Ratio_{t-1} - Ratio_{t-2}}{Ratio_{t-2}}$ instead. By taking the change in percentage any issues with either sentiment or attention still remaining after differencing are expected to be dealt with.

$$R_t = \alpha + \beta_a \cdot \%\Delta ATTN_{t-1} + \beta_s \cdot \%\Delta SENT_{t-1} \tag{4.15}$$

The fifth and final linear model is looking at the change in price at time t as a function of sentiment and attention at time $t-1$. By differencing price, it is expected to improve the stationarity of price and make it a suitable variable to regress on.

$$\Delta P_t = \alpha + \beta_a \cdot \Delta ATTN_{t-1} + \beta_s \cdot \Delta SENT_{t-1} \tag{4.16}$$

4.4.2 Exponential Regression Models

The second part of the analysis will look into exponential and logarithmic models to explore if there is an exponential relationship between price or returns and sentiment or attention. If there is no linear relationship between them it might indicate that either there is no connection or the connection is of another shape than linear.

5 Empirical Results

5.1 Summary Statistics

16

5.2 Attention & Sentiment Ratios

5.2.1 Daily Data

PCA Loadings For Attention				
Company	TC	NC	SVI	MTR
AAPL	-0.29	0.29*	-0.45	-0.44
AMZN	0.38*	0.38*	-0.33	-0.27
GOOGL	0.43*	0.45*	-0.30	-0.27
META	0.48*	0.49*	-0.23	-0.19
MSFT	-0.34	0.33*	-0.46	-0.42
NFLX	-0.34	0.25*	-0.41	-0.37

Table 4: * indicates that the loadings refer to the lag of the variable

PCA Loadings For Sentiment						
Company	TPC	TNC	NPC	NNC	PCR	VMP
AAPL	0.42	-0.34	0.38	0.13*	0.02*	0.18*
AMZN	-0.31	0.28	-0.32	0.26	0.07*	-0.21*
GOOGL	-0.37	0.34	-0.38	0.31	0.04	-0.26*
META	0.42	-0.30	0.41	0.09*	-0.03	0.23*
MSFT	-0.39	0.29	-0.42	-0.32*	-0.10*	-0.25*
NFLX	0.36	-0.31	0.36	0.16*	0.03*	0.24*

Table 5: * indicates that the loadings refer to the lag of the variable

5.2.2 Weekly Data

PCA Loadings For Attention				
Company	TC	NC	SVI	MTR
AAPL	0.31	0.34	-0.33*	-0.34*
AMZN	-0.32	-0.34	0.34*	0.34*
GOOGL	-0.34	0.43*	0.39*	-0.31
META	-0.38	-0.40	0.29*	0.29*
MSFT	-0.33	-0.38	0.37*	0.35*
NFLX	0.34*	-0.33	0.34*	-0.30

Table 6: * indicates that the loadings refer to the lag of the variable

PCA Loadings For Sentiment						
Company	TPC	TNC	NPC	NNC	PCR	VMP
AAPL	0.33*	0.21	0.32*	-0.26*	-0.24	-0.16*
AMZN	-0.36	0.3	-0.36	-0.16*	0.11*	-0.19*
GOOGL	-0.22*	-0.41	0.32	-0.41	-0.20*	0.17*
META	-0.35	-0.31*	-0.37	-0.28*	0.06*	-0.16*
MSFT	0.37*	0.29	0.40*	0.41	0.09	-0.19*
NFLX	-0.28	0.27	-0.29	0.25	0.10*	-0.20*

Table 7: * indicates that the loadings refer to the lag of the variable

5.2.3 Monthly Data

PCA Loadings For Attention				
Company	TC	NC	SVI	MTR
AAPL	0.36*	0.40*	-0.31	-0.33
AMZN	-0.43	-0.39	0.32*	-0.33
GOOGL	-0.41*	-0.38*	0.37	0.36
META	0.42*	0.45*	0.24*	-0.25
MSFT	-0.16	0.34*	0.41*	0.40*
NFLX	-0.31	0.34*	0.31*	-0.31

Table 8: * indicates that the loadings refer to the lag of the variable

PCA Loadings For Sentiment						
Company	TPC	TNC	NPC	NNC	PCR	VMP
AAPL	0.28*	0.31	-0.24	0.28	-0.18*	-0.32*
AMZN	0.29	-0.21	0.27	-0.23	-0.07	0.27*
GOOGL	-0.37*	-0.28	-0.32*	-0.25	0.25*	0.18*
META	-0.31	0.24	-0.31	0.26	-0.06*	-0.28*
MSFT	0.18	0.32*	0.31	0.31*	0.27	0.27*
NFLX	-0.24	0.25	-0.24	0.21	0.04*	-0.26*

Table 9: * indicates that the loadings refer to the lag of the variable

5.3 Linear Regression Results

	AMZN	AAPL	META	GOOGL	MSFT	NFLX
$\Delta P - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta P - KPSS$	Yes	No	Yes	No	No	Yes
$ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$Return - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$Return - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes

Table 10: Stationarity for daily values

	AMZN	AAPL	META	GOOGL	MSFT	NFLX
$\Delta P - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta P - KPSS$	Yes	No	Yes	No	No	Yes
$ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$Return - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$Return - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes

Table 11: Stationarity for weekly values

	AMZN	AAPL	META	GOOGL	MSFT	NFLX
$\Delta P - ADF$	Yes	Yes	Yes	No	No	Yes
$\Delta P - KPSS$	Yes	No	Yes	No	No	Yes
$ATTN - ADF$	Yes	No	Yes	Yes	Yes	No
$ATTN - KPSS$	Yes	Yes	No	Yes	Yes	Yes
$SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\Delta SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$Return - ADF$	Yes	Yes	Yes	No	Yes	Yes
$Return - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta ATTN - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta ATTN - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta SENT - ADF$	Yes	Yes	Yes	Yes	Yes	Yes
$\%\Delta SENT - KPSS$	Yes	Yes	Yes	Yes	Yes	Yes

Table 12: Stationarity for monthly values

5.3.1 Model One

	Daily	Daily	Weekly	Weekly	Monthly	Monthly
Proxy	ATTN	SENT	ATTN	SENT	ATTN	SENT
AAPL	-0.86	-0.48	0.82	-0.46	13.62	2.68
	(-1.6938)	(-0.9454)	(0.3243)	(-0.1834)	(1.4097)	(0.2817)
AMZN	-0.19	0.14	-0.09	2.83	-4.58	0.58
	(-0.3614)	(0.2713)	(-0.0335)	(1.028)	(-0.3964)	(0.05)
GOOGL	-1.32	0.9	1.31	2.19	-5.17	15.8
	(-3.0868)	(2.1079)	(0.56)	(0.937)	(-0.6503)	(1.9876)
META	-1.05	-0.98	3.33	-6.83	13.36	11.02
	(-2.1081)	(-1.9591)	(1.2116)	(-2.4874)	(1.3547)	(1.1529)
MSFT	-2.01	1.71	-0.0	3.3	-10.66	17.86
	(-4.0277)	(3.4201)	(-0.002)	(1.66)	(-1.7263)	(2.8747)
NFLX	-1.27	-1.02	1.13	-0.62	-21.23	-17.11
	(-1.3881)	(-1.1145)	(0.1939)	(-0.1054)	(-1.0908)	(-0.8788)

Table 13: Model One results

5.3.2 Model Two

	Daily	Daily	Weekly	Weekly	Monthly	Monthly
Proxy	ATTN	SENT	ATTN	SENT	ATTN	SENT
AAPL	0.0	-0.0	-0.07	-0.01	0.67	-7.44
	(0.2005)	(-0.0983)	(-0.6692)	(-0.1895)	(1.1331)	(-2.1652)
AMZN	-0.0	0.01	-0.12	-0.07	-0.04	0.7
	(-0.6265)	(0.6147)	(-0.8808)	(-0.4769)	(-0.3689)	(0.7771)
GOOGL	0.0	0.0	0.0	0.02	-0.47	0.59
	(0.422)	(0.3093)	(0.0744)	(1.1639)	(-0.3762)	(0.5194)
META	0.02	0.01	-0.06	0.02	0.09	-0.11
	(1.1709)	(1.408)	(-0.9325)	(0.2015)	(0.3995)	(-0.8145)
MSFT	0.01	-0.01	0.02	-0.01	-1.01	-1.53
	(0.3206)	(-1.0772)	(0.8772)	(-0.0516)	(-1.7856)	(-1.3231)
NFLX	-0.0	0.01	-0.28	-0.11	-0.26	0.01
	(-0.2135)	(0.5797)	(-0.7071)	(-0.5334)	(-0.0912)	(0.1351)

Table 14: Model Two results

5.3.3 Model Three

	Daily	Daily	Weekly	Weekly	Monthly	Monthly
Proxy	ATTN	SENT	ATTN	SENT	ATTN	SENT
AAPL	-0.25	0.03	-0.94	-1.58	3.21	3.92
	(-0.7743)	(0.0891)	(-0.6114)	(-1.0557)	(0.5563)	(0.6889)
AMZN	-0.1	0.01	1.41	0.35	-7.5	-0.66
	(-0.3076)	(0.0397)	(0.8124)	(0.2047)	(-1.0456)	(-0.096)
GOOGL	-0.71	0.45	0.6	2.06	-1.06	10.6
	(-2.615)	(1.6579)	(0.4201)	(1.4394)	(-0.2256)	(2.3491)
META	-0.81	-0.58	0.38	-2.05	-0.81	0.05
	(-2.4593)	(-1.7728)	(0.2176)	(-1.1894)	(-0.0998)	(0.0067)
MSFT	-1.36	0.94	1.13	0.45	-8.56	9.92
	(-4.3162)	(3.185)	(0.9415)	(0.3808)	(-2.0766)	(2.7855)
NFLX	-1.37	-1.06	-3.25	2.11	-4.75	2.66
	(-2.3707)	(-1.9355)	(-0.9473)	(0.6196)	(-0.4148)	(0.2272)

Table 15: Model Three results

5.4 Orthogonalizing Monthly Data

5.5 Results From Exponential Models

5.6 Comparing to Fama-French

6 Discussion

7 Conclusion

References

[1] Krishnan, Gautam Gopal. *Continued Fractions*(2016). Cornell University.
https://pi.math.cornell.edu/~gautam/ContinuedFractions.pdf

[2] Murray, L Lindberg,K. *GoogleTrends*(2023) Copenhagen Business School
https://github.com/karlolofvincentlindberg/MasterThesis/blob/cf0447d141d4d69e6ebb00392473a517e757d1cb/GoogleTrends.py

[3] Murray, L Lindberg,K. *MainDataBase*(2023) Copenhagen Business School
https://github.com/karlolofvincentlindberg/MasterThesis/blob/cf0447d141d4d69e6ebb00392473a517e757d1cb/MainDataBase.py

[4] Durbin, Michael. *All About Frequency Trading* (2010). p 25-

[5] Barber, Brad M Odean, Terrance *The Behavior of Individual Investors* (2013).Handbook of the Economics of Finance, in: G.M. Constantinides & M. Harris & R. M. Stulz (ed.), Handbook of the Economics of Finance, volume 2, chapter 0, pages 1533-1570, Elsevier

[6] Rasheed, Rayhaan *Why Does Stationarity Matter in Time Series Analysis?* (2020). https://towardsdatascience.com/why-does-stationarity-matter-in-time-series-analysis-e2fb7be74454#:~:text=Stationarity%20is%20an%20important%20concept,is%20independent%20of%20one%20another.

[7] Statsmodels *Why Does Stationarity Matter in Time Series Analysis?* (2020). https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html

[8] Jaadi Zakaria *A Step-by-Step Explanation of Principal Component Analysis* (2020). https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[9] Linda Smith *A tutorial on Principal Components Analysis* (2002). http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf

[10] Holland, S m *PRINCIPAL COMPONENTS ANALYSIS (PCA)* (2019). http://strata.uga.edu/8370/handouts/pcaTutorial.pdf

[11] Jaadi Zakaria *When and Why to Standardize Your Data* (2020). https://builtin.com/data-science/when-and-why-standardize-your-data

First

AAPL	Proxy	Obs.	Mean	StdDev.	Min	Max	Skewness	Kurtosis
Daily	Return	1744	0.0012	0.0182	-0.1286	0.1198	-0.0868	6.3515
	VMP	1744	0.2692	0.0859	0.1351	0.8752	1.9144	6.4907
	PCR	1744	1.068	0.1256	0.6876	1.3676	-0.359	-0.3149
	MTR	1744	0.0069	0.0032	0.0022	0.0284	1.937	5.4037
	SVI	1744	67.9697	42.2448	22.5806	581.3392	3.3959	21.9075
	TC	1744	4394.621	5834.909	0.0	142267.0	10.1416	195.4196
	TPC	1744	234.676	506.663	0.0	7780.0	6.9968	67.6646
	TNC	1744	255.7924	604.7332	0.0	12621.0	9.7866	147.2361
	NC	1744	2605.996	1323.2332	0	15849	3.973	26.9579
	NPC	1744	36.5831	70.1706	0	900	5.1852	41.6873
	NNC	1744	32.4627	57.456	0	1226	9.4101	147.2947
Weekly	Return	361	0.0057	0.038	-0.1753	0.1473	-0.1714	2.5688
	VMP	361	0.2665	0.0843	0.1389	0.7706	1.7401	5.2031
	PCR	361	1.0599	0.1248	0.7203	1.3356	-0.2941	-0.4142
	MTR	361	0.0333	0.0133	0.0133	0.0921	1.452	2.896
	SVI	361	328.184	177.4889	131.1603	1769.0029	2.8843	14.217
	TC	361	21213.4349	20392.0752	2946	206854	3.3418	20.6436
	TPC	361	1131.3269	1686.141	24	11260	3.091	11.3744
	TNC	361	1234.8283	2071.206	49	23778	5.6562	47.6192
	NC	361	12582.5983	4210.5029	5346	33744	1.8529	5.4051
	NPC	361	176.3019	258.3003	0	1704	2.7641	9.1871
	NNC	361	156.6981	178.1693	0	2114	5.5412	48.6115
Monthly	Return	83	0.0252	0.0815	-0.184	0.2144	-0.1766	-0.3248
	VMP	83	0.2725	0.0811	0.156	0.5457	1.3255	1.802
	PCR	83	1.0618	0.1219	0.7291	1.2866	-0.3239	-0.3276
	MTR	83	0.1432	0.0454	0.0691	0.3571	1.6616	5.4883
	SVI	83	1427.091	684.5412	713.9745	4531.5231	2.3664	6.831
	TC	83	92265.6627	71651.0214	23191	392180	1.5441	3.0374
	TPC	83	4920.5904	5272.7718	229	19275	1.2154	0.5152
	TNC	83	5370.759	6216.1352	386	33989	2.2763	6.8178
	NC	83	54726.7229	11258.0573	32577	84615	0.7564	0.2093
	NPC	83	766.8072	905.72	0	4364	2.1367	4.1037
	NNC	83	681.5422	475.7299	0	3593	3.2339	16.7784

Table 16: Summary Statistics for data on Apple

AAPL	Proxy	Obs.	Mean	StdDev.	Min	Max	Skewness	Kurtosis
Daily	Return	1744	0.0015	0.019	-0.0792	0.1413	0.701	7.2301
	VMP	1744	0.3003	0.0897	0.1475	0.6724	0.8353	0.3058
	PCR	1744	0.7848	0.1192	0.5607	1.2087	0.7265	-0.0508
	MTR	1744	0.0086	0.0044	0.0018	0.0514	2.6491	13.0338
	SVI	1744	173.1209	117.6464	16.9436	813.9376	1.2916	2.3271
	TC	1744	2765.988	2423.4841	0.0	23116.0	2.3785	8.0605
	TPC	1744	145.6537	293.1171	0.0	5637.0	10.8685	156.4762
	TNC	1744	127.4719	257.7008	0.0	4899.0	9.7825	130.4275
	NC	1744	736.004	437.1935	0	4459	2.1098	7.9125
	NPC	1744	19.4467	35.5468	0	433	6.3066	50.0226
	NNC	1744	13.0023	20.242	0	200	4.5032	27.2256
Weekly	Return	361	0.007	0.0383	-0.1446	0.1852	0.1925	2.7478
	VMP	361	0.2942	0.0866	0.1478	0.5571	0.7739	0.0161
	PCR	361	0.7899	0.1226	0.5695	1.2087	0.805	0.1388
	MTR	361	0.0412	0.0165	0.0174	0.1076	1.3078	1.7812
	SVI	361	836.0136	488.1212	118.8513	2623.9173	0.6236	0.0807
	TC	361	13349.9612	9550.0784	1835	60933	1.6901	3.1546
	TPC	361	701.1773	723.1914	89	6017	3.7223	18.6366
	TNC	361	613.4155	685.1987	48	5313	3.7688	17.5563
	NC	361	3551.8698	1526.4935	1124	9773	1.0846	1.1263
	NPC	361	93.1994	96.7033	0	584	2.308	5.9829
	NNC	361	62.7645	61.8275	0	475	2.722	10.284
Monthly	Return	83	0.0306	0.082	-0.2022	0.2689	0.4181	1.0823
	VMP	83	0.2943	0.0854	0.1668	0.5147	0.633	-0.4591
	PCR	83	0.7834	0.1162	0.5881	1.1083	0.6717	-0.2081
	MTR	83	0.1793	0.0536	0.1043	0.372	1.3182	1.7214
	SVI	83	3635.7457	1985.7357	571.6205	8667.4603	0.3162	-0.6098
	TC	83	58064.2892	37526.0187	10450	177456	1.4503	1.5822
	TPC	83	3049.6988	1897.1182	835	9229	1.251	1.1863
	TNC	83	2667.988	2164.4365	358	14237	2.8038	10.8296
	NC	83	15448.494	5292.2118	6000	29145	0.5881	-0.5524
	NPC	83	405.3614	282.8439	0	1463	1.3325	2.2542
	NNC	83	272.988	168.45	0	838	1.2768	1.571

Table 17: Summary Statistics for data on Amazon

GOOGL	Proxy	Obs.	Mean	StdDev.	Min	Max	Skewness	Kurtosis
Daily	Return	1744	0.0011	0.0167	-0.1163	0.1626	0.4351	9.9369
	VMP	1744	0.2464	0.0728	0.1265	0.7304	1.4137	4.3081
	PCR	1744	0.9525	0.0978	0.6936	1.206	-0.1331	-0.5905
	MTR	1744	0.0061	0.0031	0.0015	0.0446	3.081	19.2994
	SVI	1744	72.9649	44.0074	29.1514	328.8889	2.2088	5.3244
	TC	1744	3597.1778	3135.9737	0.0	40774.0	4.4111	33.6768
	TPC	1744	104.7856	265.4023	0.0	6979.0	14.2917	301.0697
	TNC	1744	147.9444	257.8471	0.0	4489.0	7.665	85.7088
	NC	1744	1185.6921	593.4613	0	5010	2.2538	7.9893
	NPC	1744	11.4415	31.4611	0	435	6.0922	49.0371
	NNC	1744	24.5224	53.4752	0	1284	10.4626	193.194
Weekly	Return	361	0.0053	0.035	-0.1203	0.258	1.0286	8.1009
	VMP	361	0.2428	0.0714	0.1265	0.6015	1.1816	2.5431
	PCR	361	0.9478	0.0977	0.7183	1.1714	-0.0467	-0.6268
	MTR	361	0.0296	0.0117	0.0117	0.1009	1.8817	6.126
	SVI	361	352.22	105.4913	203.0563	597.6111	0.6051	-0.7447
	TC	361	17365.3158	12073.5861	2707	88182	2.3347	7.6467
	TPC	361	503.964	733.4414	72	9240	6.3729	60.9911
	TNC	361	712.7673	804.9055	65	7288	4.2653	24.0496
	NC	361	5721.3075	1884.8823	1907	12582	1.0611	1.6027
	NPC	361	55.0249	111.6916	0	839	3.5728	15.1114
	NNC	361	118.3463	164.1079	0	1769	4.5095	32.7939
Monthly	Return	83	0.0225	0.0651	-0.1324	0.2175	0.2757	0.2668
	VMP	83	0.2459	0.0714	0.1265	0.4723	0.7501	0.2255
	PCR	83	0.9538	0.0993	0.7396	1.1417	-0.1901	-0.5643
	MTR	83	5.8938	1.7639	3.3917	12.2573	1.3452	1.8302
	SVI	83	1530.9849	450.7151	957.0761	2579.3889	0.634	-0.776
	TC	83	75528.6627	42914.6372	22805	221857	1.386	1.4051
	TPC	83	2191.9398	1779.2565	578	10929	2.2602	6.7855
	TNC	83	3100.1084	1847.1523	1059	9208	1.4471	1.568
	NC	83	24884.241	4717.7396	16541	36911	0.5247	-0.1706
	NPC	83	239.3253	374.2783	0	1876	2.8662	8.1909
	NNC	83	514.7349	440.13	0	2790	2.0363	7.9428

Table 18: Summary Statistics for data on Google

META	Proxy	Obs.	Mean	StdDev.	Min	Max	Skewness	Kurtosis
Daily	Return	1744	0.001	0.0199	-0.1896	0.1552	-0.2898	10.8466
	VMP	1744	0.3025	0.0868	0.151	0.8302	1.0976	2.6699
	PCR	1744	0.7197	0.0853	0.5084	0.9667	0.4119	-0.3085
	MTR	1744	0.0093	0.0055	0.0025	0.0705	3.2606	19.0808
	SVI	1744	155.1548	96.1282	70.4394	1425.6342	4.2658	37.5793
	TC	1744	3493.5533	3932.7514	0.0	46434.0	4.2383	27.7072
	TPC	1744	123.8062	364.2583	0.0	7501.0	12.395	195.1717
	TNC	1744	169.8509	371.45	0.0	5631.0	7.9421	83.567
	NC	1744	738.9232	664.6028	0	9187	5.1697	44.0949
	NPC	1744	12.5579	29.1669	0	379	6.3089	52.4659
	NNC	1744	46.8108	109.3346	0	1535	7.997	80.4979
Weekly	Return	361	0.005	0.0398	-0.159	0.2014	-0.3415	3.7606
	VMP	361	0.2983	0.0858	0.151	0.7023	0.9458	1.4572
	PCR	361	0.7205	0.0807	0.5384	0.952	0.443	-0.228
	MTR	361	0.0448	0.0212	0.0159	0.1878	2.0853	7.7113
	SVI	361	749.294	219.0729	496.8071	2829.9739	5.4952	42.4976
	TC	361	16866.3878	15519.501	2523	111435	2.813	10.8035
	TPC	361	597.2715	1058.9199	61	10375	5.7966	40.4392
	TNC	361	820.1357	1389.0086	83	18783	7.727	83.9197
	NC	361	3564.482	2512.2803	778	24516	4.0785	25.2968
	NPC	361	60.3407	93.6114	0	657	3.0528	11.5713
	NNC	361	226.072	418.9864	0	4768	6.9952	61.8504
Monthly	Return	83	0.0208	0.0757	-0.1334	0.2716	0.5514	0.855
	VMP	83	0.2996	0.0769	0.1724	0.5421	0.7607	0.3604
	PCR	83	0.7188	0.0834	0.5605	0.9325	0.4451	-0.0659
	MTR	83	0.1949	0.0649	0.094	0.4131	1.0809	1.225
	SVI	83	3258.1151	776.5282	2454.4	7445.7516	3.2329	13.8921
	TC	83	73358.6265	53627.888	20105	260300	1.4606	1.8581
	TPC	83	2597.7711	2504.5654	415	11954	2.1373	4.6554
	TNC	83	3567.0964	4057.6748	792	30649	4.5524	26.2487
	NC	83	15503.3494	8129.4429	7063	54099	2.4818	8.4281
	NPC	83	262.4458	248.3125	0	1222	1.6121	2.5711
	NNC	83	983.2771	1283.6708	0	8224	4.1047	19.9946

Table 19: Summary Statistics for data on Facebook

MSFT	Proxy	Obs.	Mean	StdDev.	Min	Max	Skewness	Kurtosis
Daily	Return	1744	0.0013	0.0169	-0.1474	0.1422	0.2038	10.5318
	VMP	1744	0.2359	0.077	0.1263	0.8579	2.2242	9.9342
	PCR	1744	0.8583	0.1205	0.5237	1.3058	0.3701	0.4368
	MTR	1744	0.0038	0.0018	0.001	0.017	2.4406	9.7308
	SVI	1744	193.3886	155.776	30.0	1486.7777	1.9272	6.3773
	TC	1744	2050.1909	2418.5894	0.0	59136.0	9.4672	187.0733
	TPC	1744	102.2345	589.2366	0.0	22332.0	31.9392	1172.2095
	TNC	1744	80.3842	290.7095	0.0	6962.0	14.1348	255.4007
	NC	1744	1540.5195	593.575	0	7539	2.1371	12.223
	NPC	1744	18.012	38.1547	0	371	4.0733	22.3208
	NNC	1744	7.672	14.7431	0	206	6.4963	62.9225
Weekly	Return	361	0.0064	0.0309	-0.1352	0.1503	-0.0132	3.0774
	VMP	361	0.2319	0.0753	0.1264	0.6714	1.8232	5.62
	PCR	361	0.8604	0.1219	0.5438	1.2778	0.4068	0.455
	MTR	361	0.0185	0.007	0.007	0.0552	1.7294	4.7926
	SVI	361	934.0534	698.5697	208.8571	3888.6867	1.4501	2.3161
	TC	361	9895.2825	8924.8615	938	102692	4.0289	32.95
	TPC	361	493.2964	1685.204	50	30980	16.6381	299.6931
	TNC	361	388.1939	822.6045	25	10178	7.6263	73.0034
	NC	361	7436.6205	2113.3596	2446	23301	1.6505	9.1556
	NPC	361	86.8199	138.1915	0	776	2.1481	4.3122
	NNC	361	37.0305	43.5641	0	298	2.6471	9.6887
Monthly	Return	83	0.0276	0.0601	-0.097	0.1964	0.3643	0.5066
	VMP	83	0.234	0.0744	0.1264	0.5281	1.4956	3.4236
	PCR	83	0.8537	0.1204	0.5621	1.2322	0.4227	0.5835
	MTR	83	0.0804	0.0233	0.0486	0.2113	2.4221	11.0256
	SVI	83	4061.9334	2917.2892	1041.7143	14582.2278	1.26	1.4183
	TC	83	43038.5181	30907.8159	11023	167846	1.3325	1.9951
	TPC	83	2145.5422	3612.8533	506	32915	7.7559	65.9929
	TNC	83	1688.4096	2073.1041	354	13111	3.7708	16.7969
	NC	83	32344.8193	6434.9085	20015	61372	1.4024	4.4926
	NPC	83	377.6145	521.1162	0	2498	2.2027	4.4072
	NNC	83	161.0602	128.0323	0	648	1.368	1.7254

Table 20: Summary Statistics for data on Microsoft

NFLX	Proxy	Obs.	Mean	StdDev.	Min	Max	Skewness	Kurtosis
Daily	Return	1744	0.0016	0.0256	-0.1313	0.1903	0.6548	7.7286
	VMP	1744	0.403	0.1215	0.2135	0.9527	0.9961	1.0497
	PCR	1744	1.1496	0.1316	0.7653	1.529	0.1359	-0.1125
	MTR	1744	0.0222	0.0185	0.0026	0.2463	3.5016	22.9777
	SVI	1744	88.0784	65.5666	11.3333	796.0454	3.2924	18.8721
	TC	1744	1320.9966	1541.0717	0.0	15700.0	4.0284	23.5526
	TPC	1744	104.3108	283.6734	0.0	4934.0	9.8088	119.3384
	TNC	1744	85.9358	206.8672	0.0	3938.0	10.249	145.4874
	NC	1744	558.8205	435.5247	0	3534	2.2611	7.8549
	NPC	1744	12.4065	32.8205	0	495	6.903	59.8348
	NNC	1744	10.172	25.5974	0	343	7.3612	67.0234
Weekly	Return	1758	0.0016	0.0261	-0.1602	0.2573	2.288	26.9451
	VMP	361	0.3987	0.1185	0.2135	0.8431	0.8454	0.4118
	PCR	361	1.137	0.1292	0.7653	1.51	0.171	-0.0683
	MTR	361	0.1066	0.0758	0.02	0.651	2.287	9.0459
	SVI	361	425.4261	265.5097	89.8333	2044.4608	2.304	8.2749
	TC	361	6379.41	5458.2064	763.0	32248.0	1.8962	4.3607
	TPC	361	503.8172	831.0295	39.0	6917.0	4.6447	24.9974
	TNC	361	415.0859	628.4111	29.0	6174.0	5.0169	33.5053
	NC	361	2698.7756	1637.8113	631.0	9298.0	1.1098	1.3594
	NPC	361	59.9003	104.9298	0.0	826.0	3.8698	16.987
	NNC	361	49.1385	83.19	0.0	731.0	4.8589	29.5512
Monthly	Return	83	0.0328	0.1065	-0.1971	0.4081	0.7082	1.7413
	VMP	83	0.3963	0.1178	0.2259	0.8089	0.8317	0.69
	PCR	83	1.1422	0.1327	0.7807	1.51	0.2991	0.259
	MTR	83	0.4609	0.2639	0.1233	1.1926	1.1197	0.7296
	SVI	83	1850.2748	934.2201	477.5	4717.7485	1.2489	1.528
	TC	83	27746.5904	17829.4832	4688	94247	0.9283	1.1551
	TPC	83	2191.3012	2063.9004	253	10611	2.0822	4.5661
	TNC	83	1805.3735	1540.0971	184	7709	1.781	3.285
	NC	83	11738.0482	5914.7775	3459	26836	0.5004	-0.7143
	NPC	83	260.5301	263.8644	0	1368	1.5658	2.6963
	NNC	83	213.7229	201.0061	0	965	2.0451	4.6452

Table 21: Summary Statistics for data on Netflix