Limited Attention and Disagreement in Finance Social-Media Platform

By

Gabriel Wong

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy at Cardiff University

Department of Economics at Cardiff Business School, Cardiff University

December 2021

Acknowledgments

Ph.D. research is a challenging journey; I wouldn't make it through without any help. I want to express my utmost gratitude to my primary supervisor, Dr. Woon Wong, for supervising the thesis and encouraging me throughout the Ph.D. His experience and expertise in Financial Economics have helped shape and define research direction during my study. He also provided comfort when I encounter unforeseen circumstances like interpersonal relationships. I considered him as my second father.

Secondly, I am grateful to my second supervisor, Dr. Samuli Leppälä, for his assistance in developing research ideas and spending time with me. He helped me by providing feedback on drafts and correct the mistakes that I overlooked. He is also very responsible and checks in with me from time to time to see if I am okay with my studies.

Thirdly, I would like to show my gratitude to my third supervisor, Dr. Jack Li, who envisions the use of data science in economics and finance research. He introduced me to the world of the R-Programming language, and as such, it opens the door for more exciting research. As computer enthusiasts, we both enjoyed conversations about computer technology and machine learning. Indeed, part of the idea of using Word2Vec, a Natural Language Processing technique, to analyse tweets was from him, and I am very indebted for that.

Lastly, I would like to thank Dr. Woon Sau Leung for the opportunity as co-author. Although we were in different departments, I approached him, and he quickly saw my research value and potential as a co-author. He connected me to a few well-known researchers from their fields, leading to several ongoing big projects. As a Ph.D. student, this is an eye-opening experience, and I learned how research collaboration works during the process.

Limited Attention and Disagreement in Finance Social-Media Platform

Gabriel Wong

ABSTRACT

This thesis examines retail investors' ability to predict return with behavioural constraints such as limited attention and disagreement through a unique dataset from StockTwits, a finance social media platform. To quantify StockTwits tweets into sentiment values, I borrowed from the computer science literature using the state-of-the-art machine learning algorithm, Word2Vec, to learn and classify tweets into three categories {*Bullish, Neutral, Bearish*}. With the rich user and stock level information from StockTwits, it was found that StockTwits sentiment predicted positively and significantly future stock returns. More importantly, such sentiment predictability decreased as the number of stocks users follow increased. This is in line with the limited attention explanation that users with more complex portfolios are inferior in assimilating information due to time constraints. I also explored a long-standing puzzle in asset pricing literature where the high risk does not deliver a higher return as proxied by CAPM beta. Hong and Sraer (2016) suggest uncertainties in fundamentals generate investor disagreement as the reason for the observed anomaly. Individuals who feel pessimistic are constrained from short selling, leading to lower future returns for these stocks. I formally test this hypothesis by using aggregate disagreement calculated from StockTwits, and confirms such findings.

Contents

1. Introduction

Over the last 30 years, we see a massive paradigm shift from traditional finance theory to behavioural finance. Traditional finance theory, i.e., CAPM by Sharpe (1964) and Lintner (1965), provides a simple and straightforward asset-pricing model. Risk and return trade-off must be positively related; one should not be able to predict a stock's return other than the model's measure of risk~~iness~~, beta. Yet when empirically testing stock data, a "zoo" of seemingly unrelated factors has been discovered to have predictive power both at time-series and cross-section dimensions. Many of these can be attributed to the model's unrealistic assumption; for example, it assumes investors act homogeneously, rationally, risk-averse, and frictionless to disseminate information.

Given a "zoo" of factors that traditional finance models cannot explain, research turns their attention by relaxing the traditional model's assumption through investor psychology. Hence, behavioural finance seeks to understand the bias that arises from irrational investors and the limits to arbitrage to correct mispricing back to fundamental values. Perhaps the best empirical data to understand this behavioural bias is investor-level trading data, such as their portfolio, order flow, the timing of execution and trade size, etc. However, these data are often highly regulated and hard to come by; hence prior studies focus on the stock analyst behaviour. These analysts are typically employed in investment banks or research firms to provide predictions or guidance about a company. Their work consists of reading company reports and filing, talking to management to give estimates in the form of, e.g., earnings expectations, price target, etc.

Several analysts' behavioural biases have been tested. For example, Clement and Tse (2005) looked at the accuracy of earning estimate prediction under the influence of over-confidence and herding. Lai and Teo (2008) examine home bias where local analysts predict earnings more accurately than foreign analysts. Diether, Malloy, and Scherbina (2002) found stocks with higher disagreement between analyst earnings estimates underperform. Analysts' behavioural biases are important because it affects institutional investors' decision as portfolio managers are the primary user who relies on analysts' forecasts for buy-sell decision, leading to mispricing of stocks.

On the other hand, Investors have become increasingly connected in today's information era. The popularization of social media has made sharing of information among investors ever easier.

Since the late-2000s, a notable recent innovation has been the emergence of social network platforms designed specifically for individuals interested in stock investments. Such platforms resemble the mainstream social networks, such as Facebook or Twitter, and allow their registered users to create and share short messages with others. These platforms quickly gained popularity among global investing communities, and many investors have spent considerable time tweeting and recommending stocks and trading ideas to others.

The increasing size and attention on these social networks have cast wonder about whether messages on such platforms contain value-relevant information. For instance, in January 2021, a group of retail investors used the social media platform “Reddit” to coordinate trades on a company called GameStop. Hedge fund arbitrageur ~~felt~~ pessimistic about the company placed uncover short-sell orders on the stock and its related derivatives, and suffered a loss from the buying pressure of retail investors, some of which went bankrupt.

In this thesis, I offer a fresh new perspective from the angle of a retail investor by understanding behavioral biases through their online social-media messages. Users of social media platforms are unsophisticated for two reasons. First, investors of these platforms are less professional because the barrier to access and the ability to analyze financial reports are usually high. Second, most investors in these platforms are amateur, self-taught investors as professional analysts like institutional investors have other means to communicate, such as Bloomberg's Enterprise IB (EIB) chat room with fellow professionals.

The social media messages were collected from StockTwits, one of the oldest and most widely used social network platforms on stock trading, from 2010 to 2017. Since its establishment in 2008, StockTwits has gained significant popularity among the global investing communities, secured a large active user base, and is recognized by the finance communities. Its tweets have been employed by professional financial-service firms, such as Bloomberg, to analyse market sentiment data and other news feeds as an alternative data source. In this study, the current generation of Natural Language Processing machine-learning techniques has been employed to estimate the sentiment of messages posted by all registered users. The final sample consists of more than 156,000 registered users on StockTwits, covering 8,575 U.S. publicly traded firms.

The first research objective explored individual investors' limited attention by leveraging the uniqueness of the tweets dataset at user-level ("investor" or "user" used interchangeably).

Limited attention is the theory that if one allocates cognitive resources across tasks, attention spent on one task must reduce attention available for other tasks. An investor's characteristic is the degree of complexity of a social network user's sets of stocks he covers. In spirit to the analyst limited attention literature, a user's degree of attention can be proxied by the groups of stocks he covered, as measured by the number of stocks or industries simultaneously.

Research consumes considerable time and effort, but those who spend resources gathering information are compensated (Grossman and Stiglitz, 1980). In the context of stock investment, researching stocks typically involves gathering company information, analysing and interpreting qualitative and numerical data, and exercising subjective judgment in formulating views and predictions about fundamental values and future prices (Jung et al., 2012; Harford et al., 2018). Due to the complex nature of such processes, the effect of limited attention on the efficacy of stock analysis by social network users is somewhat ambiguous and under-explored.

On the one hand, users with market-wide knowledge who cover large sets of stocks argue they have an informational advantage over those with a relatively less complex portfolio. As the number of stocks followed/covering increases, one cannot maintain the same level of time and effort on research as the number of stocks followed/covering increases. On the contrary, human beings have limited cognitive resources to apply to different tasks, and, hence, the attention spent on one task necessarily reduces attention available for other tasks (Kahneman, 1973). These attention constraints are particularly evident among individual investors (compared to finance professionals or institutions) who have less trading experience and are less financially sophisticated (Barber and Odean, 2013). Thus, equity research by social network users with a complex portfolio may be inferior because the time and effort allocated to each followed stock are likely to be reduced. This implies that company information may not be fully gathered, processed, and synthesized (Boni and Womack, 2006; Bradley et al., 2016). In the same vein, users following only a few stocks likely pay more attention to each and thus have more in-depth and potentially superior knowledge (Clement, 1999; Gilson et al., 2001; Corwin and Coughenour, 2008).

Alternatively, one may also argue that users following only a handful of stocks (i.e., with less complex portfolios) can be more susceptible to particular behavioural bias. It develops myopic and overly-optimistic views of companies they have spent considerable time and attention on, and with which familiarity and personal attachment may have developed (Piotroski

and Roulstone, 2004; Chan and Hameed, 2006). Whether portfolio complexity enhances or reduces the quality of equity research for social network users is ultimately an empirical question.

My regression result finds that social-media sentiment positively and significantly predicts future monthly stock returns, controlling for a wide array of firm characteristics and user and month fixed effects. The result complements prior studies (see, e.g., Renault, 2017; Bartov et al., 2018). Notably, the tests reveal that such positive predictability decreases significantly with the number of stocks users have been following over the previous six months, consistent with a limited attention behavioural explanation. Economically, estimates show that when the number of stocks covered by users is at the 25th and 75th percentiles, a one-standard-deviation increase in social-media sentiment is associated with increases in future stock returns of 37.8 and 17.0 basis points per month, respectively.

Based on these findings, it is interesting to see whether users with a complex portfolio might allocate less time and attention to analysing the fundamental position of their covered companies, whereby these users' social media tweets are less likely to predict future earnings performance than those with less complex portfolios. In line with this conjecture, it was found in my research that when the number of stocks covered by users over the past six months were at the 25th, and 75th percentiles, a one-standard-deviation increase in social media sentiment was associated with $0.008 and $0.001 higher earnings per share, respectively. The results continued to hold when earnings surprises were examined.

Furthermore, the next question is whether the predictability of social media sentiment over future stock returns depends on company characteristics that capture how users comprehend their financial positions. If attention constraints give rise to our findings, the negative moderating effect of stock coverage is likely to be more pronounced among informationally opaque companies, have complex organizational structures, and are harder to analyse. The loss in accuracy for stock-return predictions among such stocks is likely more severe when research effort and effectiveness are reduced due to users' attention and time constraints. My results from the subsample tests supported this conjecture. Overall, results from the additional tests supported the limited attention theory.

Ultimately it is a matter of whether a trading strategy can be established, hence I proposed by exploiting the return predictability of social media sentiment and testing the

profitable strategy. To explore this, the StockTwits users were divided into low and high stock coverage groups; the low group consisting of users covering 100 stocks or below, the high group containing users with more than 100 stocks. The social-media sentiment was equal-averaged for each of the two user groups and the net (low-minus-high) sentiment was computed for each stock. The trading strategy that longs a portfolio of stocks with zero and positive net sentiment, and shorts a portfolio of stocks with negative net sentiment, generates a significant monthly excess return of 21.7 basis points and a Carhart (1997) trading alpha of 26.4 basis points per month.

My finding was that social-media sentiment predicted positive and significant future returns, and importantly, such positive predictability decreased when the number of stocks users follow increased. The return predictability appeared to stem from users' ability to forecast future earnings. Further tests revealed that reduced predictability due to the limited attention in higher stock coverage was significant only for complex, opaque firms and thus hard to analyse. Consistent with limited attention literature, individual investors who covered many stocks with limit information processing due to time constraints lead to lower return prediction.

More importantly, my empirical research confirms the theoretical findings of Bossaerts et al. (2020). Their paper studies individual computational complexity where assets uncertainty is the consequence of not knowing which information is pertinent rather than not having the information in the first place. They formalize this as a 0-1 Knapsack problem, a combinatorial optimization problem whereby asking the decision-maker to find the subset of items of different values and weights that maximize the total value knapsack subject to a weight constraint. In the laboratory experiment, individuals endowed with several securities trade to maximize this knapsack problem. The result is consistent with my finding where price information generally revealed incorrection solutions, and the informational quality deteriorated as ~~the instance~~ complexity increased.

My second research objective studied how investors' dispersion of beliefs with social media tweets disagreement affected the pricing in the cross-section of stocks return. The first and most persistent anomaly was found empirically that the high beta stocks earned lower return than the low beta stocks, which is against the theoretical prediction of the Capital Asset Pricing Model (CAPM) by Sharpe (1964), Lintner (1965), and Mossin (1966). Baker, Bradley, and Wurgler (2011) suggest since January 1968, portfolios sorted with higher beta exhibit a cumulative

decline in return. For instance, a dollar invested in a value-weighted portfolio of the lowest quintile of beta stocks will have yielded $96.21 ($15.35 in real terms) at the end of December 2010, while a dollar invested in the highest quintile of beta stocks would have yielded around $26.39 ($4.21 in real terms).

Early explanations for the flatness of the empirical Securities Market Line (SML) was given by Black (1972), who suggest that all investors are short-sale constrained and cannot borrow at the risk-free rate, resulting in a flatter SML line compared to the unconstrained ones. Black's proposed model is accurate to the individual investor in the real world but, to some extent, not for the institutional investor as they can borrow much more freely, just not infinite amount as assumed in CAPM. Indeed, hedge funds with 1.8 trillion dollars of assets under management can short-sell. Only those retail mutual funds that have more than 20 trillion dollars of assets under management are prohibited from short-selling directly (Almazan et al., 2004) or indirectly through the use of margins and derivatives (Koski and Pontiff, 1999). This idea is then further developed by Frazzini and Pedersen (2014). They modelled investors who face limits on their leverage, and the slope of the SML line depends on the current market condition of the tightness of the leverage constraint. In other words, relaxing leverage constraint results in SML line that is more consistent with CAPM.

Hong and Sraer (2016) offer a novel way to look at the flatness of the SML line; they argue that high beta assets are speculative and offset the risk-sharing motive of investors. They show that relaxing the CAPM central assumption of homogeneous expectations and costless short-selling results in a downward sloping or invert-U shape SML line. In their model, investors disagree on the mean value of a common market factor. High beta asset naturally demands higher loading in common market factor as investors naturally disagree more on the cash flows during a highly uncertain macro-economy. In simple terms, beta amplifies disagreement about the common market factor.

The model also prohibits some investors from short-selling where retail mutual funds (MFs) cannot short while hedge funds (HFs) can. The high disagreement promotes different trading of high beta assets by different investors; with short-sell constraints, the pessimists are sidelined from short selling them while optimists acquire more. This creates overpricing of the high beta assets.

Several empirical evidence supports the idea of investor disagreement; for instance, household and professional forecaster shows time-series variation of disagreement about many of the macro-economic variables such as industrial production (IP) growth, market earnings, and inflation; see Cukierman and Wachtel (1979), Kandel and Pearson (1995), Mankiw, Reis, and Wolfers (2004), Lamont (2002). These aggregated disagreements come from well-documented behavioural biases such as heterogeneous priors or cognitive biases like mental accounting, overconfidence, etc. See Barber and Odean (2001), Lamont (2004).

Miller (1977) and Yu (2011) hypothesized and found where disagreement could come from a portfolio of stocks, ~~and~~ a market portfolio would mimic macro-economic disagreement. For example, Park (2005) was among the first to measure top-down market disagreement using analyst forecast dispersion of S&P 500 index for annual earnings-per-share (EPS) to proxy for macro-economic disagreement. However, it is not straightforward to understand how analysts base their EPS forecast on a portfolio of assets. On the other hand, a portfolio disagreement constructed from the bottom-up by aggregating disagreements with individual stocks in a portfolio has two advantages. First, the bottom-up measure favours a better signal-to-noise ratio than the top-down measure. Second, bottom-up aggregate disagreement is created from thousands of stocks with stocks covered by many analysts, whereas the top-down approach involves fewer analysts.

In this study, I used social media tweets to proxy for retail investors, currently underexplored in the literature, to create a bottom-up measure of aggregate disagreement at the stock level. Social media tweets related to stocks can be classified as both “bullish” and “bearish” directions. Analogous to optimists and pessimists, I first constructed a stock-level disagreement measure and follow Hong and Sraer's (2016) theoretical implication where beta varies positively with stock-level disagreement to create a beta-weighted aggregate disagreement.

A beta-sorted portfolio was formed to establish how beta varies with stock-level disagreement from social media tweets and returns from 2010 to 2017. The beta of each stock was estimated in the cross-section using the past 12 months of daily return and adjusted for small stock illiquidity using the method by Dimson (1979). NYSE stocks’ pre-ranked beta is used to form the twenty beta-sorted to minimize the small stock effect. A portfolio’s beta was also calculated with value or equally-weighted return. Social media disagreement at the portfolio

level was also formed by averaging disagreement at the stock level. My result was in line with Hong and Sraer's (2016) theoretical prediction that social media disagreement at the stock level varied positively with portfolio betas, while future 12 months portfolio return exhibited a concave pattern. For instance, portfolios 1, 10, 20 have disagreement (returns) of 0.508 (3.32%), 0.518 (12.44%), 0.636 (7.94%) respectively.

Having understood the relationships between portfolio beta, disagreement, and returns, I constructed a beta-weighted aggregate disagreement from social media tweets to proxy market-wide macro disagreement. To understand the determinant, the source, and the prediction of aggregate social media disagreement, a regression was performed with several macroeconomic variables that could potentially explain my findings. Results intuitively suggested that a higher market Dividend-Price ratio, Industrial Production Index (a measure of real output in the economy), Consumption-Wealth ratio, and Term-Spread predicted lower aggregate social media disagreement, whereas a higher VIX, the CBOE volatility index predicted higher disagreement. Since the aggregated social-media disagreement consists of a majority of retail investors, the regression result implies that their opinions follow the macroeconomic trends and environment.

Lastly, a thorough examination of the mechanism of how aggregated social-media disagreement affected the SML line was also performed by using a Fama-Macbeth (1973) style regression test. In the test, 20 portfolios' returns were regressed onto their beta, beta-squared and their coefficients were retained each month. This was followed by a time series multivariate regression to uncover the relationships between beta, beta-squared coefficients on aggregate social-media disagreement, and various macro-economic variables as control. Notably, aggregate social-media disagreement exhibited positive (negative) and significant relationships with beta (beta-squared) independent variables at different lengths of return horizon, implying a concave SML line.

The overall evidence established that retail investors from finance social-media platforms were unlikely to be irrational noise traders. Various macro-economic variables could predict their judgment in aggregate disagreement. As with any other investor type, retail investors who speculated on high-beta assets and were more sensitive to disagreement on these cash flows. Together with short-sell constraints where retail investors experienced the most, these stocks were over-priced. Empirical tests confirmed this finding where higher disagreement in high beta

stocks had lower future returns while low beta stocks were unaffected, giving rise to a concave SML line.

This thesis is organised as follows: Chapter 2 presents a literature review on behavioural finance and social media in research. Chapter 3 explains the primary data – StockTwits and the machine learning approach to measure information content within the tweets. Chapter 4 explores limited attention in the theme of StockTwits users, which is proxied by the number of stocks they covered and subsequently the predictivity of returns. Chapter 5 understands the social media user's disagreement in the stock market and their implication to the stock return as a whole. Finally, chapter 6 concludes.

2. Literature Review

This chapter provides an overall literature review from the failure of traditional finance theory to the emergence of behavioural finance, placing more weight on limited attention and disagreement. In the second and third sections, I outlined the use of social media in recent research and calculated social-media sentiment using data provided by social-media platform StockTwits.

2.1 Literature Review: Behavioural Finance

The starting point of traditional finance theory dates back to Fama (1970) influential paper titled "Efficient Capital Market: A review of theory and empirical work". He outlines the backbone of testing market efficiency, namely the Efficient Market Hypothesis. The null hypothesis outlines that the market is efficient and stock prices fully reflect all available information; any arbitrage opportunity is impossible. As a result, stocks trade at fair value, making investors impossible to buy undervalued (overvalued) stocks. It is also important to stress that it is impossible to outperform the market by picking stocks and timed trading opportunities. In other words, the only way to gain extra return is to purchase the higher-risk asset. Fama (1970) also developed three forms of information efficiency by relaxing the null hypothesis into weak, semi-strong, and strong efficiency, setting the groundwork for research.

Under the weak form efficiency, stock price fully reflects all information under historical prices and volume only. For semi-strong form efficiency, stock price fully reflects those in weak form, together with any publicly available information. Publicly available information includes any news, company, or analyst's reports. Lastly, strong form efficiency requires all information reflected in stock price from the prior two efficiencies. Any private information, that is to say, any insider news will be included in stock price.

An influential paper rejects the idea of a perfectly informationally efficient market. Grossman and Stiglitz (1980) provided a paradox and argued that information will always be costly, price won't perfectly reflect all information quickly. If it did, no one would be willing to spend resources to gather such information and, in return, receives no compensation. There is simply no reason to trade in the market that would collapse. Therefore at equilibrium, the market must contain a degree of inefficiency such that investor is willing to expend, gather, trade and profit from obtaining this information.

During the same period, Sharpe (1964) and Linter (1965) developed the Capital Asset Pricing Model (CAPM), which is an elegant, easy to interpret model to capture the idea of pricing assets with the relationships between the risk and expected return. The model contains four main assumptions. The first one assumes the market is perfect with no friction, meaning that trading or information gathering comes immediately and secondly it does not incur transaction costs or taxes, borrowing, or lending at risk-free. Third, it requires investors to have a homogeneous expectation, maximizing own's utility subject to budget constraint, acting rationally, and being risk-averse. Lastly, CAPM assumes investors only hold assets for one period, and individuals cannot influence the asset's price (price taker).

The CAPM model has been tested by various scholars, Black, Jensen, and Scholes (1972), who used 35 years of monthly data and found that, on average, low-beta stocks outperformed high-beta stocks. That is, the empirical SML line is flatter than what CAPM's SML line predicted. In other words, the low beta asset has positive alpha while high beta stocks have negative alpha. Baker, Bradley, and Wurgler (2011) showed that if one invested $1 into a low quintile beta-sorted and value-weighted portfolio in January 1968 yields $96.21 ($15.35 in real terms) at the end of December 2010. This empirical test shows the unsatisfying performance is attributable to the unrealistic and overly simplified assumption by Hong and Sraer (2016).

Over the 1970s, numerous "anomalies" in finance have been discovered, and researchers failed to use traditional risk-based models (such as CAPM) as explanations. The list below summarizes the anomalies during this time:

1. **Size effect** refers to the fact that more prominent firms have lower returns than small firms. On average, Banz (1981) suggested small firms receive 0.40% higher monthly returns than large market capitalization stocks. Some theories suggest that small firms have a more significant growth opportunity than larger firms and tend to be more volatile and sensitive to good news, leading to an enormous price appreciation. Furthermore, small firms have a lower stock price, and it attracts a more casual retail investor to enter the market, creating a volatile stock price.
2. **Value effect** relates to firms trading at a lower price compared to their potential fundamental values. Investors try to capture such inefficiency by searching, e.g., Accounting ratios of high Earnings to Price (E/P) and book-to-market (BtM) provide a higher average return than low ratio firms. Although Basu (1977) was the first to

document such anomaly, he also shows more unexplained residual other than market factor as in CAPM.

3. **The momentum effect** first documented by Jegadeesh & Titman (1993) found a positive correlation between today's and yesterday's price such that a trend exists. They formed momentum strategies using data from 1965 to 1989 and reveals that buy past winner (high return over past 3 to 12 months) and short sell past losers (low returns over past 3 to 12 months) by holding for future 3 to 12 months earns monthly returns of 1% for next year. The momentum effect gives rise to technical analysis, where such an effect is considered an oscillator and is used to identify trends. In short-term trading, one would follow the trend by selling low, buying lower, or buying high and selling higher instead of a rather monotonic buy-low, sell-high strategy.
4. **The January effect** is a perceived seasonal increase in stock price during January. Two main explanations exist; Berges, McConnell, and Schlarbaum (1984) suggested a tax-loss selling hypothesis whereby selling assets when the price has gone down in December helps reduce year-end tax liability. The second hypothesis relates to institutional investors' tendency to rebalance their portfolios yearly by buying winner and sell losers to show a better alpha.
5. **The weekend effect**, first observed by Cross (1973), is significantly lower stock returns on Monday than those on the preceding Friday. Some theories explain a tendency for companies to release bad news on Friday after the market closed, leading to a fall in price on Monday. Others suggest it might be related to short-selling, particularly stocks with high short interest positions. Such effect was studied by numerous academics and found to have disappeared from 1987 to 1998. Though starting from 1998, the effect reappears again.

The anomalies above give rise to the use of human psychology to explain such phenomena. Behavioural finance focuses on the following pillars:

Informational processing is the error in information processing that leads the investor to misestimate the true probabilities of associated returns. In general, there are four such errors.

1. **The forecasting error** suggested by Kahneman and Tversky (1972, 1973) found that people place too much weight on recent experience compared to prior beliefs when

making forecasts. As a result, they tend to make too bold or herd forecasts along with their peers, giving inaccurate and inaccurate uncertainty in their information.

2. **Overconfidence** is where people often overestimate their skills and capabilities. Some investors substantially diversify their portfolio in investment, too confident in investing in multiple asset classes without conducting thorough research. For example, Malmendier and Tate (2004) found that overconfident CEOs create corporate investment distortion. These CEOs overestimate return from investment and underweight external funding as costly. Hence, they overinvest when the internal fund is abundant but curtail investment when requiring external funding.
3. **Conservatism** gives rise to slow information diffusion; investors are too slow in updating their beliefs in the face of new evidence, giving rise to momentum in stock market return. Therefore, conservatism bias can be considered as one source of underreaction.
4. **Sample size neglect and representativeness** refer to people who generalize information using a subset of samples and believe this represents the whole sample. Therefore, they infer patterns too quickly based on a small sample and extrapolate trends too far into the future, leading to overreaction.

Behavioural bias is where investors often make consistent or systematically suboptimal, irrational decisions. Ritter (2003) listed out some of the critical behavioural biases:

1. **Framing** refers to how an idea or term is exhibited to people. In other words, decisions are affected by choices. For instance, investors will act risk-averse in terms of gains but risk-seeking in terms of losses. The research found that people react differently and choose the latter in these two scenarios, although the outcomes are the same: The first scenario is a coin toss with a payoff of $10. A gift of $10 that bundles with a bet that imposes a loss of $10. In both cases, the probability of winning or losing is 50%.
2. **Mental Accounting** is a specific form of framing where people segregate particular decisions. Simply put, people tend to split decisions that wouldn't require splitting. For example, a gambler would gamble more if they are ahead. In the same vein, after some consecutive success, one may think the total bet is funding out from the pool of gains he has previously, making him more tolerant of risk.

3. **Regret avoidance** is where people tend to regret more when the decision made was unconventional. For example, a loss in blue-chip stock will not be as painful as a loss in a start-up firm that you are not familiar with.
4. **Disposition effect/Prospect theory** tends to sell assets that have increased in value while keeping assets that have dropped in value. For example, an investor purchases a stock at $20; the price drops to $12 and rebound to $16 afterward. Psychologically, people won't sell the stock unless it goes above $20, and as such, investors are risk-averse in small gains but risk-loving for the losses. Kahneman and Tversky (1979) then developed this idea called Prospect theory. In their theory, when individuals are given two equal choices, one has possible gains and others with potential losses, individuals will choose the former choice even though both would yield the same economic result. I.e., Receiving $50 would be the same as gaining $100 then lose to $50. However, both will receive $50; individuals would prefer the former scenario. Their theory also directly challenges standard utility theory, where utility assumes individuals have only one risk appetite measured at absolute terms. Still, prospect theory suggests one can have a different risk appetite depending on gain or loss.

Lastly, **Limits to arbitrage** are one of the most important pillars in behavioural finance. Logically, behavioural biases would not cause anomalies if rational arbitrageur can fully exploit the mistakes made by the behavioural constrained investor. For example, if rational arbitrageur saw a behavioural investor pushing price up, he would correct such mispricing back to fundamental value. However, in the real world, Barberis and Thaler (2003) listed several factors that hinder such correction from limiting the ability to profit from mispricing.

1. **Fundamental risk** occurs when there is a potential rational revaluation as new information arrives. Suppose a firm had a very negative earnings announcement and the stock price currently falls below the stock's fundamental value. Although a rational arbitrageur may take this opportunity to profit by buying the stock, doing so is not risk-free because if there is further bad news in the future, ordering to buy the stock may pose a loss. Thus, one would need to re-evaluate the fundamental value.
2. **Noise-trader risk** refers to investors who don't base their trade on fundamentals and follow buy-sell decisions solely on price rules. Such an effect would push the price away from fundamental value even more and for a long time. For example, suppose a

stock that is currently underpriced. One could buy the stock to correct the mispricing. However, this is not risk-free because the noise-trader can order excess selling pressure. In any case, the price would go back to fundamental value, though not in the short term. The arbitrageur would require a long investment horizon to counteract any short term fluctuation in price.

3. **Implementation cost** is related to the difficulty of exploiting overpricing. For the retail investors who wish to arbitrage overpricing, one needs to short-sell the stock. Short-selling is when the investor anticipates a fall in stock price, one will need to sell the stock at the current price and buy back in the future. Without owning any stocks, he would need to borrow the stock from others today and return it to the lender later. Two critical difficulties present: he needs to pay lender interest, and shorting will not be successful if the lender requires the stock back on short notice.

2.2 Recent Development: Behavioural Finance

From the previous section, it is not hard to see why anomalies still exist even arbitrageur made an effort to push prices back to fundamental values. Hong and Stein (1999) proposed a model that incorporated and simplified all the behavioural biases humans exhibit and explained why the price wouldn't trade at fair value, giving rise to under(overreaction) of stock price. The model setup is as follow: Suppose there are two groups of investors called news-watchers and momentum trader; both groups are not entirely rational and are boundedly rational. That is, they are only able to receive and process a subset of available public information. The model assumes news-watchers make their buy-sell decision based on currently available company fundamentals and do not use price in their judgment. However, news-watchers require time to process, spread around their population, and absorb news leading to the lagged buy-sell decision. On the other hand, momentum trader uses a straightforward heuristic rule of thumb based on past prices. One can think of them as trend followers where a simple decision to buy if the price increase.

Focusing on only one stock, in the first period of their multiperiod model, it spawns with a group of the news-watchers trader where positive company information appears. Such information would have pushed the stock price up immediately to the new fundamental value in a frictionless market. However, their model contains friction where price adjusts slowly to new information. This is because information diffuses gradually, and some news-watchers require time to process the news. News-watcher who processed the information quickly purchase the stock. Still, all news-watcher do not use stock price as their information set; slow news-watcher underreacts, leading to a sluggish stock price adjustment to the long-run fundamental value.

In the second period, suppose there is no other news about the stock. Having seen the slight price increases, momentum traders join the market, pushing the stock price even more. The price increase in the second period sets off even more momentum traders in the market until the price overshoots the long-run fundamental value leading to overreaction. News-watchers realize the price is too high in several future periods and started their downward price adjustment; momentum traders follow suit to create trend reversal. Understandably, early momentum traders earn higher returns than those late joiners. They also do not worry about losing money because such a trend-chasing strategy, on average, makes money.

Limited attention

Having understood the news-watcher investors type under Hong and Stein (1999) model, some academics begs what makes gradual information flows and why individuals take time to process information leading to the lagged decision. This idea formed the basis of limited attention. In fact, Kahneman in 1973 already proposed the idea of attention limit and argued humans have "limited attention" in a sense if one allocates cognitive resources across tasks, attention spent on one task must reduce attention available for other tasks. Hirshleifer and Teoh (2003) and Peng and Xiong (2006) later stressed the idea that cognitively overloaded investors pay attention only to a subset of available information.

In Hirshleifer and Teoh (2003) paper, they develop a theoretical model to bridge the gap between the traditional and behavioural view of asset pricing, where “inattentive” investor creates “mispricing” in equilibrium. The model follows three ingredients: First, part of the investors behaved like unsophisticated retail ones; Second, the valuation error by unsophisticated investors are systematic; Third, limit to arbitrage constrain the arbitrageur or sophisticated investor from taking a position to eliminate the mispricing. The market consists of two exogenously determined investors, which means they cannot choose to become. Limited attention investors represent fraction *f* of the market, and attentive investors represent *(1-f)* of the fraction in the market.

These attentive investors are rational because they are using full information signals to form investment decisions. Though some uncertainty exists on the firm's cash flow, no other signals in the economy could provide them with this additional information. On the other hand, limited attention investors do not receive full information signals. This group of investors believed they knew everything. They will not look for additional information sources to update their beliefs; they also consist of those who receive zero signals, which model them as cognitively biased or overloaded. The interpretation of the model of limited attention investors is that they are fully rational investors with high information cost to obtain signals. Or the fact that they have limited information processing capabilities.

The model develops as a single-period stock market model with perfect information disclosed initially and paying a liquidating dividend at the end. The equilibrium is a Walrasian equilibrium, meaning that all investors are price takers who cannot influence price and do not use price as an information variable. The attentive investor knows everything in the market, while

limited attention investors do not know they only know part of the information. All investors follow mean-variance preferences and must allocate their wealth into one risky asset and a riskless asset. No diversification is allowed meaning that the model requires pricing of idiosyncratic, firm-specific risks. The objective of investors is to estimate the end-of-period value of the firm based on the information available to them. Suppose the current stock price is P_1 and end-of-period stock price with uncertainty is P_2, investors demand x number of shares according to this formula:

$$x = \frac{E(P_2|I) - P_1}{A\,Var(P_2|I)}$$

Where E is the expected value given information I at the start of the period, A is the risk aversion coefficient, and Var is the variance of the end-of-period cash flow.

This formula suggests investors will demand higher if the firm expects to be worth more at the end-of-period, while it is a decreasing function when he requires to pay to buy the stock now and is governed by the risk he likes to bear with the risk aversion coefficient. All types of investors calculate the expected value of the price of the stock. However, limited attention investors assumed to calculate the same wrong value because they received the same limited information and traded aggressively on their firm's inaccurate valuation. In contrast, attentive investors trade according to their correct expected value.

At the equilibrium, the aggregate demand and supply of shares is the weighted average of the expected value of the two groups of investors less the discount of risk. The stock price is biased to its intrinsic value because of the existence of limited attention investors. Hence, the misvaluation depends on the bias size made by limited attention investors and their population size in the model. Moreover, all investors do not believe the price is right in a sense where if limited investors overvalue the stock, they will buy more than if the firm is at intrinsic value. Attentive investors will purchase fewer shares; however, their action does not undo mispricing. The reason is that attentive investors will still hold the stocks if they expect the future return in positive, mispricing prevails.

Empirically, Corwin and Coughenour (2008) empirically tested the limited attention hypothesis using individual NYSE specialist portfolios and tried to understand how liquidity provision is affected as specialists allocate their attention across stocks. Specialists, also known as market makers, provide an ideal setting to test because they are obligated to provide liquidity

for a set of securities. As a result, one can directly identify the set of securities across which the specialist must divide his attention. Their research question asks if limited attention forces a specialist to allocate effort across stocks. His ability to provide liquidity for a given stock will be negatively related to the attention required for other stocks in his portfolio. A successful trade requires matching bid and ask order, and a specialist's job is to step in should there be a buy-sell imbalance on the bid, ask order, and such imbalance can be seen in the price movement and the bid-ask spread of the stock. Using TAQ (Trade and Quote) intraday transaction panel data, their main result includes a firm fixed-effect model to remove the firm's time-invariant heterogeneity with specification:

$$Liquidity\ Measure_{it} = \alpha + \sum_{p=2}^{13} \alpha_i \times HH_p + \beta_1 InvPrice_{it} + \beta_2 LogTrades_{it} + \beta_3 LogTradeSize_{it} + \beta_4 AbsReturn_{it} + \sum_{j=1}^{J} \gamma_j PanelAttention_{jit} + \epsilon_{it}$$

Where $Liquidity\ Measure$ includes $\{Rate\ Of\ Price\ Improvement,$ $Dollar\ and\ percentage\ magnitude\ of\ price\ improvement, Percentage\ effective\ Spread\}$ of a stock i at time interval t, $InvPrice_{it}$ is the inverse of average trade price during period t, $LogTrades_{it}$ is the natural log of one plus number of trades during period t, and $LogTradeSize_{it}$ is the natural log of one plus the average trade size during period t, $AbsReturn_{it}$ is the absolute midpoint-to-midpoint return during period t. $PanelAttention_{jit}$ contains three versions of attention. For example, the first attention is the standardized trade frequency of stock i at time t. A dummy variable HH_p is also included to identify 13 half-hour trading periods from 9:00am to 4:00pm to control for intraday patterns in transaction cost.

Their finding shows that the rate and magnitude of price improvement decrease and bid-ask spread increase as specialist's attention to other stocks increases. In other words, when a specialist focuses more stocks in their portfolio, the liquidity would reduce for each stock in his portfolio.

Sonney (2009) analysed financial analysts' sector vs. country-wide specialization in their stock coverage to their earnings forecast precision. To examine this, he obtained earning forecast from the I/B/E/S dataset, and the specialization of an analyst is measured using Herfindahl Index

(HI) with $HI_{ay}^{Country} = \sum_{c=1}^{C} \left(\frac{N_{cay}}{Nay}\right)^2$ and $HI_{ay}^{Sector} = \sum_{s=1}^{S} \left(\frac{N_{say}}{Nay}\right)^2$ for country and sector specialization. Where N_{cay} and N_{say} denotes analyst a number of firm coverage in country c or sector s at fiscal-year y; N_{ay} is the total number of firms coverage for analyst a at fiscal-year y. To gauge financial analysts' performance, he used percentage demeaned absolute forecast accuracy, which he called $PDAFA_{a,j,t,y} = -\frac{DAFE_{a,j,t,y}}{\overline{AFE_{j,y}}}$. Where $\overline{AFE_{j,y}}$ is mean absolute forecast error calculated over all analysts' forecasts of firm j's earnings in fiscal year y. $DAFE_{a,j,t,y} = AFE_{a,j,t,y} - \overline{AFE_{j,y}}$ is the demeaned absolute forecast error and $AFE_{aj,t,y} = \left|EPS_{j,y} - F_{a,j,t,y}\right|$. The $PDAFA$ adjust for example the biased forecast error due to analysts' herding. The main result uses the following specification:

$$PDAFA_{a,j,t,y} = \alpha_{j,y} + \beta_{ABS}ABS_{a,y} + \beta_{COS}COS_{a,y} + \beta_{SES}SES_{a,y} + \sum_{l=1}^{L} \gamma_l Z_{a,j,l,t,y} + \epsilon_{a,j,t,y}$$

Where $ABS_{a,y}$ is a dummy variable equals to one if analyst a is an "absolute" specialist and zero otherwise, $COS_{a.y}$ is a dummy variable equals to one if analyst a is a country-specialized analyst over fiscal year y and zero otherwise, $SES_{a,y}$ is a dummy variable equal to one if analyst a is a sector-specialized analyst and zero otherwise. $Z_{a,j,l,t,y}$ are a vector of control variables.

Although firm-year fixed-effect is used in the regression, some concerns not adding analyst fixed-effect would bias the estimate as the analyst's skills or ability are central. To control for this, he added analyst controls to capture this effect. The main result suggests that analysts specializing in country coverage provide a more accurate earning forecast than their sector-specialized peers. In other test, comparing specialization within country and sectors, he found a higher number of countries (sectors) coverage reduces accurate forecast. One explanation is that there is limited attention when covering many countries (sectors) because analysts' stock portfolio becomes more complex. Though, comparing a country with sectors, analyst develops a myopic view with sector specialization not knowing market-wide information renders them underperform.

In stock return anomalies literature, Dellavigna and Pollet (2009) posit the question of "Does limited attention among investors affect stock returns?". They examine the decision where attention to new information plays a crucial role in response to earnings surprises by comparing announcements that occurs just before the weekend, on Friday, to announcements on other

weekdays. The hypothesis is that humans have time constraints and limited cognitive resources to process information; however, no prior research understands limited attention when the investor is distracted. If the weekend distracts investors and lowers their quality of decision-making, the response to Friday earnings surprises should be less pronounced. Investors revise their decisions in subsequent periods, and the information must eventually incorporate into stock prices. They measure this delay response using post-earnings announcement drift with month indicator, that is, month fixed-effect, to control for differences in return sensitivity across quarters and within a quarter (early vs. late announcement). They found that Friday announcements have 15% lower immediate response and 70% higher delayed response.

Da, Engelberg, and Gao (2011) proposed a direct measure of investor attention using the Google search volume indicator (SVI). SVI is weekly constructed using aggregate search frequency to measure attention by the search term for each stock's ticker, company name, and company abbreviation. They use a sample of Russell 3000 stocks from 2004 to 2008 and find that SVI correlates with other investor attention proxies, such as news media coverage. The measure also captures timelier than any other methods using the Vector Auto Regression model. In terms of stock price prediction, they found higher stock prices with increasing SVI in the next two weeks and a trend reversal within a year, confirming prior studies such as Hong and Stein (1999) model.

Baker and Wurgler (2006) also study how investor sentiment as a proxy for attention affects the cross-section of stock returns. Using various proxies for the sentiment from 1962 to 2001 and form a composite index using principal component analysis based on the following six series: Close-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. They found that when the starting period sentiment is low, the following returns will be high for small, young, highly volatile, unprofitable stocks, etc. Conversely, when sentiment was started high, the stocks earned lower subsequent returns.

Disagreement

Literature in disagreement tries to understand investors under how heterogeneous prior affect differences in belief investors hold. In this strand of literature, academics ask the question of what motivates people to trade? The hypothesis is that it requires that investors disagree on company fundamental or stock price such that trading could happen. If everyone agrees, the stock price will not move until additional new information arrives. A concrete example illustrates such an idea. Suppose that a firm announces a 5% increase in earnings this quarter. For the first type of investor who expects constant growth in earnings and believes that this earning increase is permanent, one will stay neutral about the view on the present value of expected future earnings, which is 5%. The second type of investor who also expects a constant increase in earnings but believes such an increase is short-term would view the constant growth optimistically and as a piece of positive news. The third type of investor expects a 15% increase in earnings and experience disappointment in the 5% growth and revise expected future earnings downwards.

Although three types of investors all receive the same news, they are inclined to trade with others due to their heterogeneous prior beliefs. As such, disagreement highly correlates with trading volume. Many previous researchers also found such trading volumes correlate positively to returns. For instance, glamour stocks, especially in the 2000s dot-com bubble, are overpriced. As a result, these glamour stocks generate a much higher trading volume than value, lower-priced stocks. Tracing this anomaly, Basu (1977) first found where glamour stocks with a high P/E ratio earn lower returns than value, low P/E ratio stocks.

Noted that disagreement in this context generates trading volume, it is insufficient to explain why glamour is overvalued (higher priced) than value stocks. Indeed, disagreement is related to short-sale constraints; investors often sit out of the market when they value the stock as overvalued since short-selling is nearly impossible. For instance, Almazan, Brown, Carlson, and Chapman (2004) use a sample of U.S. domestic equity funds from 1994 to 2000 to understand mutual fund managers' restrictions on trading. They found that investor and securities commission restricts fund managers in borrowing, purchasing securities on margin, holding individual equity options, trading in equity index futures, or purchasing restricted securities. In particular, merely 11% of funds were allowed and was proceeded to short sell in 1994.

Lamont and Stein (2004) literature review paper document the relationships between short interest and stock returns using NYSE-Amex and Nasdaq stocks from 1960 to 2002. The rise and fall of stock prices during the dot-com bubble was due to a surge of interest in short-selling and found that short-selling is undertaken mainly by rational arbitrageurs, for example, hedge funds. And the demand for short positions is highest with overvalued, high price to book ratio stocks. Their second finding suggests there are frictions in the market like institutional rigidities, paying premiums for the duration of shorting, uncertainty in price movement making stock borrowing very costly. In effect, stocks that require short-selling the most tend to have abnormally low future returns.

Theoretically, Miller (1977) presents a static model and points out that when investors disagree about a stock's future value, the optimist will buy, and the pessimist will sell. Thus, in a frictionless market, prices reflect the average opinion. However, with short-selling constraints in place, the divergence of opinions creates stock overpricing; pessimists cannot short-sell when they do not hold a physical stock, while optimists can buy and sell freely. Consequentially, the optimists will set the price, and the pessimist becomes the price taker, overpricing the stock.

This "overvaluation hypothesis" lends support from Duffie, Garleanu, and Pedersen (2002). In particular, they provided a framework for the price impact from shorting. Given that short-selling is time-consuming, the authors analyse the price effect of agents' differences in opinions from trading and how lending fees are determined and factored into the price. Their model first predicts that a significant difference in beliefs between pessimists and optimists, together with demands for short-selling and buying the stock, can have a strong enough effect to push the price way above the valuation of the most optimistic investors. Secondly, their result shows that contrary to general thinking, the price would not necessarily fall when the short-selling constraint is relaxed. Instead, a negative stub value mispricing can arise from lending fees, which is the implied stand-alone value of the parent company's assets without the subsidiary after being carved out in Initial Public Offering (IPO).

A notable example illustrates this. 3Com owned palm in the 2000s, and on March 2nd, 2000, 3Com sold a fraction of ownership to the public, initiating an IPO. In this equity carved out, 3Com retains 95% shares of Palm and spins off remaining shares to 3Com's shareholders. Thus, 3Com shareholders would gain 1.5 shares of Palm for every share of 3Com they owned. If the law of one price is valid, 3Com must be at least 1.5 times the price of Palm. The day before

Palm IPO, 3Com closed at $104.13 per share. The next day after the IPO, Palm closed at $95.06 per share, meaning that 3Com should have jumped to $145. Instead, 3Com fell to $81.81, a negative stub value of 3Com was minus $63, and the market did not adjust for another few months. Clearing Palm was overpriced. Market participants with a difference in opinion, particularly those pessimists who believe Palm's price is too high, would trade on it. After exhausting different explanations, Lamont and Thaler (2003) argue that short-sell constraint with lending fees and stock liquidity contributes to this mispricing.

In terms of empirical evidence with disagreement on stock returns, Diether, Malloy, and Scherbina (2002) used I/B/E/S analyst's dispersion with long-term EPS growth as a proxy for stock's disagreement stocks with higher disagreement earn lower future return than other stocks. They first obtain stocks from NYSE, AMEX, and Nasdaq exchange and require firms to have Compustat book equity data from the fiscal year ending in calendar year $t-1$. Their sample period starts from January 1976 to December 2000.

In their research design and filters, stocks with share prices lower than $5 are removed, and the rest are assigned to portfolios based on dispersion on analyst forecast following Jegadeesh and Titman (1993). Specifically, they assign stocks into five quintiles each month according to the analyst dispersion from the previous month, and stocks are held for one month. The portfolio return is the equal-weighted average of the return of all stocks in the portfolio.

Results suggest a portfolio of stocks in the highest quintile of disagreement underperforms a portfolio with low disagreement by 9.48 percent per year. Such effect is also more pronounced in small stocks and stocks that have performed poorly in the past year. Dische (2002) also documents a portfolio of high dispersion stocks yields 0.80% return per month while low dispersion stocks yield 1.74% higher return each month.

Serafeim and Yoon (2021) investigated disagreement in the literature on Environmental Social and Governance (ESG). In recent years more data vendors started to provide ESG performance indicators for investment managers to trade. Many of these ratings are multidimensional, hard to quantify, and post challenges to identify the ESG outcomes. For example, it is hard to quantify how a firm contributes to climate change. Their research focuses on how disagreement across different data vendors could affect a firm's stock news and market reaction. Using industry and date fixed-effect model, they analysed and found that firms with

disagreement have the moderating effect of ESG rating, which diminishes the prediction on future news. Specifically, the test has the following specification:

$$ESG\ News_{i,t} = \beta_0 + \beta_1 Average\ ESG\ Rating_{i,t-1} + \beta_2 Disagreement_{i,t-1} + \beta_3 Average\ ESG\ Rating_{i,t-1} \times Disagreement_{i,t-1} + \sum_{j=1}^{J} \beta_j Control\ Variables_{i,t-1} + \epsilon_{i,t}$$

Where $ESG\ News$ is the ESG News score from the news provider TruValue Labs, $Average\ ESG\ Rating$ and $Disagreement$ are the averages and standard deviations from the five ESG data vendors.

In particular, consistent with prior literature of disagreement on stock returns, higher disagreement between vendors weakens the prediction of future return and hence concluded that rating disagreement hinders ESG information incorporated into the price.

Finally, Bargeron, Lehn, Moeller, and Schlingemann (2014) study whether disagreement between managers and investors affects the information contained in acquirer firm's returns in the setting during mergers and acquisitions. They estimated a regression that uses year fixed-effect to control in time-invariant macroeconomic shock that could affect the M&A. Their dependent variable includes, for instance, *Bidder change volatility* is the change in implied volatility that measures the market's disagreement on the acquiring firm's stock. Control variables include both acquiring firm's CEO characteristics, deal characteristics, etc... They find that the coefficient of interest, *CAR3*, is the acquirer firm's 3-days cumulative abnormal return which acts as a proxy for investor's reaction around the deal announcement date is inversely related to its implied volatility. The result confirms the disagreement hypothesis when the market is uncertain about the valuation of acquiring firm, and the investor disagrees more with the managers of the acquiring company.

2.3 Literature Review: Social Media Sentiment

The past decade has seen an explosion of new sources of information and communication that are easily accessible to a market participant. With the introduction of Web 2.0 applications, websites based on interactive user-generated content over the last 15 years have presented many new opportunities for engagement. In addition, social media sites create communities that allow people to share ideas, collaborate on, and share content. Wikipedia is one prime example of a free online encyclopaedia that gives users full access to review, edit and create articles. With quick and easy-to-use platforms, social media presents itself as a practical tool for mass communication and gaining a phenomenal user base in recent years. The most extensive dissemination of information is the social media platforms such as Twitter, where users post a concise message to convey their ideas and view about stocks to a broad audience. As of 2018, active social media users in the U.S. amounted to 243 million, or roughly 75% of U.S. adults get their news from social media (Suciu, 2019).

Given the rich content in social media, it is the subject to study in various disciplines. For example, in marketing, Alves, Fernandes, Raposo (2016) review studies of consumers from the usage, share, and social media influence consumer decisions and perceptions. They study how customers help the firm build brand loyalty, promote products and services, brand followers, brand recognition, and brand recall. They also look from the firm perspective on the use of social media, the implementation, optimization, and measurement of results. In this case, they focus on marketing intelligence, promotions, public relations, product and customer management, and marketing communications. In criminology literature, researchers Geoff and Bell (2012) study how geographically created communities with different political ideologies, common interests from social media can play a significant role in spreading the threat, hate speech, and growth of terrorism. Tourism research has also seen growth in using social media. For instance, many aspects of tourism, especially information search and decision-making behaviours, tourism promotion, and searching for best practices for interacting with consumers via social media channels, e.g. sharing holiday experiences in social networks.

The use of social media is also very prominent in finance. Tashtego, a hedge fund firm based in Boston, set up a Social Equities Fund with trading ideas based on sentiment from social media.[2] DataMinr, a start-up firm that parses Twitter feeds to generate trading signals, suggests

[2] http://fortune.com/2015/04/02/hedge-fund-twitter/

they raised $130 million in financing.[3] Such data is undoubtedly an emerging new information source to investors, be it individual or institutional investors. These messages or tweets create crowds' wisdom that aggregates information provided by non-expert individuals could predict outcomes more precisely than experts.

In the finance literature, only a few have tested theories using social-media messages. Perhaps the earliest one by Antweiler and Frank (2004) examined more than 1.5 million messages posted on Yahoo! Finance and Raging Bull about 45 companies in Dow Jones Industrial Average and Dow Jones Internet Index. Using the number of messages posted or bullishness of these messages, they consider three main questions: whether it predicts return, volatility, and generates trading volume? They first created a bullishness index by calculating the number of bullish messages minus the number of bearish messages, scaled by a total of bullish and bearish messages. Finding suggests it helps predict volatility, though trading volume is more critical to predict volatility on some firms, and while these messages are more important than trading volume in others. They found that message posting helps predict negatively for stock return prediction, and their results are significantly small but robust. Lastly, they also tested the disagreement literature. They confirmed the model of Harris and Raviv (1993) that under two regression specifications whereby regressing stock return on current or future bullishness index, both associated with more trades.

Tetlock (2007) examine the role of media and the stock market using news from the Wall Street Journal column. Firstly, he constructed a measure of media content that corresponds to negative investor sentiment or risk aversion. His finding suggests pessimistic media content forecasts patterns of market activity. A higher value of media pessimism creates downward pressure on market prices and leads to higher trading volume. More importantly, the price impact of pessimism is significant and slow to reverse in small stocks. Exploring the predictive relationships between pessimism and negative market return provides robust support.

On the other hand, Bartov, Faurel, and Mohanram (2018) examined social media using Twitter and tested where individuals tweeted just before the firm's earnings announcement to predict earnings and announcement returns. They hypothesize whether aggregate opinion from individual tweets predicts quarterly earnings, how opinion predicts stock price reaction to firm's

[3] http://www.wsj.com/articles/tweet-analysis-firm-dataminr-raises-funding-14265664862/

earning realization, and test for the information quality of firms affect the future aggregate opinion of individual tweets. Result suggests aggregate Twitter opinion does predict the company's quarterly earnings, positive association of stock price reaction to aggregate twitter opinion. The findings are also more pronounced for firms with weak information environments, proxied by lower analyst following and lower institutional ownerships.

Giannini, Irvine, and Shu (2017) examined the home bias in behavioural literature using Twitter messages. Investors who are geographically close to the firm earn higher returns due to information advantage, as Ivkovic and Weisbenner (2005) and Seasholes and Zhu (2010) documented using discount brokerage individual investors' portfolio holding. Using Twitter message and creating a sentiment variable at every 2-weeks interval, they first examine the full sample and found negative relationships of 8.3 basis-point in abnormal stock returns in the subsequent week. This negative relationship is significantly robust from 2 days to 1-month return windows. Next, using the distance of the user's location and the firm headquarter within 100 miles defines users as local and nonlocal. Using these two groups to form Tweet sentiment, they find that nonlocal tweet sentiment exhibits more negative future returns.

Interestingly, local sentiment has no significant predictive ability and is a "nonlocal disadvantage" rather than "local advantage." A portfolio long-short strategy is used by shorting positive differences between nonlocal and local sentiment stocks. Buying a negative difference in sentiment stocks yields an abnormal profit of 6 bps per day. They conclude that nonlocal sentiment favours many "glamour" stocks that tend to be overpriced and lower future returns. These stocks also contain higher information asymmetry between firm and investor, leading to investors making consistent mistakes in predicting future returns.

Recently, Cookson and Niessner (2020) studied investor disagreement from social media. Using a new dataset from StockTwits, they can map out investors' different approaches (e.g. technical, fundamental) and examine how much disagreement is driven by different information sets. Finding suggests differences in investment philosophies drive investor disagreement. Using date fixed-effect regression model, they explore various types of disagreement, within-group and cross-group disagreement, and found that both sources of disagreement lead to more trading volume. However, within-group disagreement has 2.5 to 4 times more abnormal trading volume.

2.4 Summary for Chapter 2

This chapter reviews the literature on the transition from traditional finance theory to the emergence of behavioural finance. I discussed the foundation of the traditional finance theory - Efficient Market Hypothesis. EMH provides different forms of informational efficiency, all of which laid the groundwork for researchers to test how information incorporates into an asset's price. Based on this, empirical researchers discovered many anomalies that the traditional risk-based model cannot explain. For instance, the size effect presents where large firms have a lower return than small firms. The January effect is where January earns a higher return on average compared to other months. The weekend effect shows Monday has a lower return on average than other days in a week.

Looking for an explanation, researchers borrowed from the psychology literature and created a new strand of the field called behavioural finance. Behavioural finance argues most anomalies focus on three main biases. Firstly, information processing bias leads investors to misestimate the actual returns probabilities. Secondly, behavioural bias is where investors often make consistent or suboptimal mistakes. Lastly, limits to arbitrage are the frictions in the market obstructing rational arbitrageur to correct mispricing back to the fundamental value of an asset.

To formally understand how behavioural finance affects stock valuation, starting from the early 2000s, Hong and Stein (1999) developed a notable model explaining why the price would not trade at fair value, giving rise to under or overreaction to stock price. The model incorporates early behavioural explanation like limited information processing of investor and noise trader who follows the trend to make returns to models limits to arbitrage for rational arbitrageur.

An essential part of the model is that if an investor receives information differently, what makes the information flows gradually, and why does the investor take time to process information leading to a lagged decision? Kahneman (1973) first proposed the idea of limited attention, which states humans have time constraints. If one allocates cognitive resources across tasks, attention spent on one task must reduce attention available for others. Numerous theories and empirical studies have confirmed that limited attention is a critical behavioural bias in assets mispricing.

Another ingredient to assets mispricing is investors having a difference in opinion. For example, suppose investor receives the same information. After careful assimilation, two groups of investors arrive with a verdict of being optimistic and pessimistic about a stock. High trading volume happens when both parties disagree with each other in great magnitude. In theory, the

market should be efficient as both parties trades, and fundamental value settles at equilibrium. However, individuals' short-selling constraint, which is one of the limits to arbitrage, prohibits pessimists from short-sell leading to stock overpricing. When short-sell is not permitted, disagreement mispricing is confirmed in a seminal paper by Diether, Malloy, and Scherbina (2002). Using analyst earnings forecast dispersion as a proxy for disagreement, buying the lowest quintile disagreement portfolio and short-sell highest quintile produces a 9.48 percent return per year.

Lastly, I provided a literature review on the recent research using social media as alternative data to answer questions related to finance. Individuals are more connected with each other since the start of Web 2.0, allowing web users to create content interactively in the last 15 years. Contents created contain valuable information that interests researchers in many different fields. From marketing to understanding how social media build a firm's brand loyalty, promotions, public relations, etc., to criminology relating social media to hate speech, spreading political ideologies, etc. In particular social media in finance literature have received significant attention to understanding users' trading behaviour, especially from the wisdom of crowds. I reviewed several finance papers that examined how social media sentiment relates to or even predicts stock market returns, volatilities, earnings, and trading volume. Most studies concluded that social media is vital in examining stock market indices.

3. Data and Method

3.1 Introduction

In this thesis, my main goal is to examine limited attention and disagreement from the viewpoint of individual investors. Therefore, this chapter will introduce a new data set from StockTwits, a social-media platform for stock discussion, and lets the readers understand the dataset's structure used in the subsequent chapters and explains how to extract information from the tweets. I will then explain how to remove noises from the data to reduce measurement error and the machine learning technique used in measuring each tweet's sentiment. Lastly, a discussion of the appropriate econometric methods are provided in modelling the StockTwits dataset.

3.2 StockTwits

The primary data used in the following chapters comes from a company called "StockTwits". StockTwits was founded in 2008 as a social-media platform that allows investors to share their opinion about stocks. According to Ian Rosen, CEO of StockTwits, it is one of the oldest and most popular platforms. There are about 350,000 monthly active users, and more than 150,000 messages exchanged every day. Unlike general social-network platforms such as Twitter, where users can tweet about any topic, StockTwits users mainly post tweets about stock trading, often sharing trading ideas and stock-price predictions/recommendations. These tweet messages are labelled using "cashtags" to signify the stock referring to (e.g., $TSLA). Tweets are concise, short, and cannot exceed 160 words. While users tweet about any topic related to stock trading, tweets are grouped into two types. Firstly, it focuses on the fundamental position of companies, for example, payout policies, asset composition, financial ratios, earnings, filings, etc. An example of such tweets is shown in Figure 1a, where a user shared a bearish opinion about Boeing Company and lowered the price target due to slowed production rates and safety issues related to their 737-MAX planes and the risk of China-US trade war. Another example is present in Figure 1b. where a user provides a "bullish" sentiment for Tesla Inc. based on its newly opened factory in China. Another type of tweet message focuses on technical analysis where users often post tweets containing technical indicators, such as candle stocks, support and resistance levels, moving averages, etc. For instance, Figure 1c shows the user suggested a short strategy for Apple Inc. if its stock price was around $270. In Figure 1d, a bearish prediction that the stock price of Amazon.com, Inc. would fall below the current point of support.

StockTwits has several functions that facilitate interaction between users. For example, registered users can express their agreement with a post by giving it a "like", commenting on it, or even sharing it in their profile. Users can "follow" other users to automatically get updates and future tweets from the latter. Furthermore, StockTwits also allows users to label tweet messages as "Bullish" or "Bearish", a unique feature that makes sharing views among users more friendly and convenient.

Figure 1
Examples of Tweets Posted by StockTwits Users

Fig. 1 presents four examples of tweets posted by StockTwits users. Fig. 1a and 1b (1c and 1d) show a user who posted tweets relating to fundamental (technical) analysis. In Fig. 1a, the user showed a bearish sentiment for The Boeing Company because of the ongoing 737-MAX flight issues and forecasted reduced production. In Fig. 1b, the user signalled a bullish sentiment for Tesla, Inc based on its newly opened factory in China, signalling for growth. In Fig. 1c, the user suggested a short strategy for Apple Inc. if its stock price were around $270. In Fig. 1d, the user shared a bearish view on Amazon.com, Inc. after its stock price fell below the support.

Fig. 1a

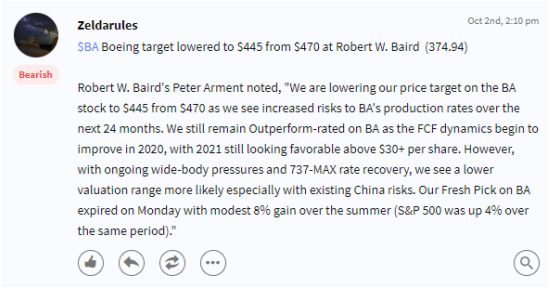


Fig. 1b

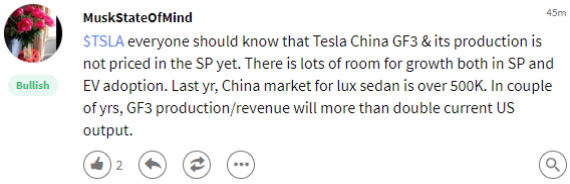


Fig. 1c

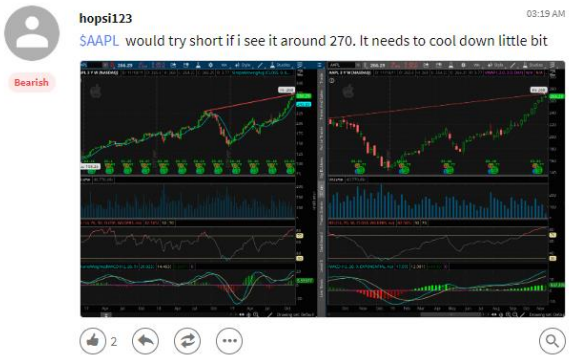


Fig. 1d

Driven by these useful features and relative ease in exchanging trade ideas, StockTwits has rapidly gained popularity. In 2010 it was acknowledged as one of the 50 best companies by The Times Magazine. In 2012 it was recognized as one of the top 10 most innovative finance companies by Fast Company Magazine, a leading American publication on technology and business. Roof (2016) points out that over 60% of active StockTwits users were young professionals aged under 44. Apart from a large community of active users, the StockTwits

platform is also used by professional financial services providers, such as Bloomberg and Google, to track and estimate market sentiment data and generate a news feed.[4]

I downloaded all tweets (more than 48 million) posted by more than 280,000 unique registered users on StockTwits over January 2010 to the end of December 2017 and collected text content, company tickers, and the ID of users who tweeted for each tweet. User-level characteristics such as other users following, or being followed by, a user and the number of likes received from other users were also downloaded. Since our goal is to analyse stock-return predictability, any tweets without a valid company ticker or "cashtag" are removed.

There are concerns where the data selection confines to StockTwits, for example, many other platforms such as Facebook, Twitter, or other big data sources like Google Trend and other specialized content (e.g., Seeking Alpha). Indeed other platforms can provide valuable information on the stock discussion. However, social media like Twitter, Facebook are not specialized in finance. These tweets or messages will contain more noise compared to a finance-specialized platform. Secondly, big data sources like Google Trend only provide aggregated data, for example, search trends about a topic. Although Seeking Alpha is a specialized content provider, much of the information is concluded from a user pool. StockTwits provides tweets and information down to the user's level, as mentioned above, which allows for interesting behavioural analysis.

[4] For instance, Bloomberg had a Bloomberg Social Velocity (BSV) function to track stock sentiment using StockTwits and Twitter data feed.

3.3 Identifying and removing bot users

One crucial concern in social media is that some users are "social media bots" run by computer algorithms. They often post a considerable amount of repeated and nonsense tweets. For example, "Estimize" (https://stocktwits.com/Estimize) is a finance research platform that crowdsources earnings and economic estimates from a range of hedge funds, brokerages, independent and amateur analysts. They often write tweets about companies missing or meeting EPS consensus forecast and other fundamental news indicators. While some bots do indeed post informative and economically important tweets, the inclusion of such tweets will likely create bias and measurement errors. Moreover, the objective of this study is to extend the understanding of investment-decision and test human users' behavioural biases, including such tweets and bots into our estimation cast doubt on our findings. To this end, I propose and adopt methods famous in machine-learning literature for removing bots.

Following prior machine-learning literature (Efthimion, 2018; Kearney, 2018), who highlighted two important criteria that are shown to be essential to predict bot users. Firstly, human users have limited attention, cognitive ability, and time to make unlimited posts on the platform. Users who post an excessive number of tweet messages on an average day are likely to be bots. In the same vein, users who follow excessive numbers of other users on an average day are likely to be non-human users.[5]

The approach to solving this problem is for each user over the sample period (since they have joined), I compute the number of tweets posted per day, and secondly, the average number of users followed on an average day. Since users with extreme values in the two fields are very likely to be bots, I manually checked each user in the top 1 percentile of each field (23.4 tweets per day and 60 users followed) to analyse common patterns bot-user names. In total, 37 identified keywords were suggestive of bot users should they appear in the username, including "alert", "app", “bot”, “filing”, “network”, “news”, “press”, “url”, etc. There were 6,354 users removed whose names contained any of these keywords from the raw sample, reducing users to 277, 389. While such exclusion criteria would inevitably exclude human users, this approach is conservative and hence preferable. Table 1 below shows the name list for bot users.

[5] I thank Dr. Michael Kearney for making his machine-learning program publicly available: https://github.com/mkearney/tweetbotornot

TABLE 1
Name List For Bot Users

This table gives the list of words that likely suggest bots when appearing in the names of users. I obtained this list by manually examining each user in the top 1 percent in (1) the average number of tweets posted per day (*# of tweets per day*) and (2) the average number of users followed per day (*# of users followed*). I report the number of users whose names contain the key words and the percentile statistics for both ratios.

		# of tweets per day						*# of users followed*					
List	# of users	Mean	1%	25%	Median	75%	99%	Mean	1%	25%	Median	75%	99%
Alert	253	11.552	0.000	0.245	1.234	4.500	188.275	13.156	0.000	0.000	0.143	3.517	232.000
Adviser	10	5.616	0.002	0.091	5.000	7.000	16.000	180.897	0.000	0.007	161.750	317.000	540.000
Advisor	58	2.318	0.000	0.125	0.777	1.579	48.000	6.722	0.000	0.002	0.077	0.585	217.000
Analyst	154	6.907	0.000	0.099	0.641	2.867	174.667	6.026	0.000	0.004	0.097	0.417	60.000
Analytic	52	6.735	0.000	0.079	0.282	0.853	235.750	3.883	0.000	0.008	0.055	0.468	60.000
Analyze	31	3.678	0.000	0.449	1.357	3.667	19.000	49.097	0.000	0.005	0.393	60.000	666.000
App	866	2.056	0.000	0.043	0.267	1.262	31.250	5.793	0.000	0.008	0.102	0.612	60.000
Bank	308	3.215	0.000	0.049	0.310	1.478	30.000	15.560	0.000	0.010	0.087	0.584	280.000
Bot	366	4.268	0.000	0.065	0.392	1.757	74.333	4.021	0.000	0.012	0.087	0.461	60.000
Capital	799	2.315	0.000	0.075	0.356	1.515	32.929	14.077	0.000	0.011	0.085	0.452	454.000
Chart	458	2.469	0.000	0.103	0.431	2.148	25.537	5.048	0.000	0.006	0.074	0.423	220.000
Company	24	2.866	0.000	0.020	0.087	0.483	44.500	2.867	0.000	0.027	0.108	0.502	30.000
Filing	5	84.777	0.087	0.200	0.709	19.394	403.494	0.127	0.000	0.000	0.006	0.020	0.609
Fund	336	2.249	0.000	0.061	0.500	2.446	24.667	11.834	0.000	0.011	0.095	0.809	240.000
Hedge	210	2.079	0.000	0.069	0.437	1.766	21.000	2.354	0.000	0.006	0.074	0.364	60.000
Hype	105	2.749	0.000	0.200	1.324	4.528	13.000	55.828	0.000	0.018	0.213	10.000	581.000
Insider	87	11.450	0.000	0.092	1.000	2.500	224.410	8.518	0.000	0.000	0.084	0.378	353.000
Live	428	4.055	0.000	0.070	0.500	3.000	74.200	26.406	0.000	0.014	0.118	1.366	533.000
LLC	324	2.110	0.000	0.056	0.350	1.446	37.018	4.950	0.000	0.010	0.103	0.520	60.000
Mutual	3	0.803	0.178	0.178	0.379	1.853	1.853	0.415	0.246	0.246	0.277	0.722	0.722
Network	27	13.287	0.000	0.033	0.439	2.937	295.089	3.899	0.000	0.032	0.126	0.389	60.000
News	257	9.287	0.000	0.144	1.000	3.485	204.600	6.501	0.000	0.007	0.139	1.000	60.750
Partner	34	21.864	0.000	0.056	0.319	1.072	536.000	17.634	0.000	0.004	0.122	0.367	393.000
Platform	8	14.882	0.108	0.826	2.325	10.179	92.291	11.674	0.000	0.023	0.993	16.179	59.000
Predictor	10	5.630	0.191	0.200	0.832	6.349	36.455	1.353	0.000	0.002	0.324	0.800	8.571
Press	95	5.778	0.000	0.048	0.295	1.957	183.571	15.049	0.000	0.006	0.077	0.573	653.000
Rating	17	26.918	0.000	0.290	8.400	18.695	295.089	2.026	0.000	0.024	0.113	0.654	25.000
Reporter	14	13.148	0.000	0.031	0.271	0.880	175.201	1.614	0.000	0.009	0.049	0.321	20.333
Service	20	3.382	0.000	0.012	0.142	1.750	52.159	26.558	0.000	0.005	0.162	15.298	350.000
Shop	101	2.914	0.000	0.119	0.346	1.838	47.500	13.084	0.000	0.013	0.070	0.347	358.000
Stocktwits	107	3.259	0.000	0.029	0.519	3.833	27.000	4.685	0.000	0.007	0.112	0.604	60.000
Ticker	96	5.656	0.000	0.060	0.286	1.474	322.467	8.007	0.000	0.004	0.102	0.626	289.000
Top10	7	6.363	0.090	0.585	2.533	12.000	23.000	100.135	0.000	0.000	0.193	30.000	666.667
Tracker	88	9.915	0.000	0.328	8.500	18.000	38.000	35.138	0.000	0.001	0.352	60.000	306.000
Trend	390	2.280	0.000	0.062	0.347	1.625	51.263	4.282	0.000	0.005	0.047	0.333	75.000
Url	123	2.662	0.000	0.038	0.200	1.125	34.000	4.514	0.000	0.005	0.115	0.339	60.000
Wire	83	4.699	0.000	0.041	0.200	1.781	111.000	5.268	0.000	0.011	0.124	0.562	154.000

After screening users by names, additional filters were introduced to ensure that bot users could be readily removed. Specifically, the top 0.1 percentile (92 tweets on an average day) in the number of tweets posted daily and the top 1 percentile (60 users followed on an average day) in the daily average of users followed were excluded, further removing 2,317 users.[7] Finally, the tweets were merged with stocks in CRSP and Compustat. After excluding any unmerged data and missing values in the stock and accounting variables, our final sample consisted of 156,657 users, posting slightly below 30 million tweets. Details on the sample attritions can be found in Table 2 below.

TABLE 2
Sample Attrition

Panel A of this table reports the percentile statistics for *# of tweets per day* and *# of users followed per day* for all users, and those users who are identified as bots based on names (see Table IA.1). Panel B shows how our sample is reduced due to each of the exclusion criteria. The number of users and the average *# of tweets per day* and *# of users followed per day* are reported.

Panel A. Average user characteristics

	All users						
	Obs.	Min	5%	25%	Median	95%	Max
# of tweets per day	283,538	0.000	0.000	0.032	0.202	7.040	7125.000
# of users followed per day	283,538	0.000	0.000	0.009	0.091	30.000	2000.000
	Names identified bots						
	Obs.	Min	5%	25%	Median	95%	Max
# of tweets per day	6,149	0.000	0.000	0.066	0.409	13.000	1068.754
# of users followed per day	6,149	0.000	0.000	0.008	0.094	60.000	1571.000

Panel B. Applying the exclusion criteria

		Average	
	# of users	# of tweets per day	# of users followed per day
Raw	283,538	1.803	5.000
After removing names-identified bots	277,389	1.757	4.862
After the filter on *# of tweets per day* (@99.9th)	277,112	1.540	4.847
After the filter on *# of users followed per day* (@99.0th)	274,795	1.483	3.047
After excluding tweets unmatch with CRSP	156,657	1.327	0.288

[7] Another type of user which would cause measurement errors are fake-news gangs who are humans but manipulate public sentiment by spreading fake or distorted news. This was the case when Twitter cracked down on state-backed users in 2019 (Twitter Safety, 2019). Since coordinative and manipulative behaviors by "fake-news gangs" require the ownerships of many user accounts and large number of users followed by them to be effective, the filter based on extreme number of users followed are likely effective in removing them.

3.4 Measuring social-media sentiment: A machine-learning approach

Prior finance literature has applied machine-learning techniques to classifying text (e.g., Giannini et al., 2017; Bartov et al., 2018; Cookson and Niessner, 2020). I utilized state-of-the-art technologies from Natural Language Processing (NLP) to measure social-media sentiment. More specifically, using FastText, an algorithm developed by Facebook AI Research (FAIR), to classify user posts into "Bullish" and "Bearish" tweets. [8] Tweets unrelated to stock trading are cleaned. For example, those containing only conversational dialogues, website URLs, or those written in non-English characters were removed. Moreover, I take reasonable steps to encode non-textual information into useful sentiment data. For instance, the emoji icon, 😊, is replaced by "smiling face with smiling eyes", whereas 😞 was coded as "disappointed face".

A commonly used method in measuring social-media sentiment is to generate a dictionary of keywords or synonyms for certain sentiment states (e.g., bullish and bearish) (see, e.g., Loughran and McDonald, 2011; Bartov et al., 2018). Under this approach, the classification model searches each tweet for those keywords and generates a sentiment score based on the frequency counts of matched keywords. While this approach is cost-effective and straightforward, erroneous classifications could arise. For instance, the word "bad" is considered bearish, while more complex phrases or expressions containing "bad", such as "not bad", have a completely different meaning and should be regarded as bullish. More importantly, social media tweets contain abbreviations and non-frequent use words. Using a dictionary approach may undercount, leading to a biased sentiment score.[9]

Another method more advanced than the keyword approach is 'maximum entropy', used in studies such as Giannini et al. (2017) and Cookson and Niessner (2020). This method employs a learning algorithm that counts the frequency of words appearing and models the conditional dependence of words to analyse a sentence's sentiment structure, thus enabling the correct identification of semantic information embedded in multiword phrases in a tweet.

[8] FastText have been used frequently by Facebook in chatbot and marketing ads. For example, in chatbots application it learns the sentiment from what user typed, classifies sentence into the correct category, and shows the required results. (https://www.engadget.com/2016/08/18/facebook-open-sourcing-fasttext/)

[9] I also considered using quantitative software NVivo for analyzing unstructured text like tweets. However, NVivo also relies on a predefined dictionary and counts the words for each sentiment category (Bullish, Bearish) to present a sentiment value. This method suffers from undercounting due to many abbreviations and non-frequent word use in tweets and would underestimate the sentiment value.

FastText, the algorithm employed in my study, is inherently different from the above methods. Specifically, FastText utilizes neural network architecture to treat each word and multiword phrase as a series of vectors and store their textual information as a numerical value.[10] It learns the meaning of single words or word phrases in association with their neighbouring words. An essential advantage of FastText is that, by using sub-word information, it is robust in classifying text with irregular formatting or misspelled words. For instance, users may have written "Gooood" instead of "Good" by mistake or occasionally convey a stronger feeling. Tweets may also contain typos such as "Appel" when referring to "Apple." Likewise, abbreviations such as the use of "w/o" to replace "without" are also frequently used. Through learning and modelling at the character level, FastText identifies and corrects these instances, while models from previous generations may wrongly classify these irregulars and misspelled words as separate words.

To implement the FastText algorithm and allow it to "learn," all StockTwits that were explicitly labelled by users as "bullish," "bearish," and "neutral" were used in the training sample, which was then further split into a training set (80%) and a validation set (20%), the latter used to verify the out-of-sample model's predictive accuracy. Multinomial logistic regression was employed to train the model. The predicted probabilities, fitted using the estimated parameters from the logit model, were then used to classify unlabelled tweets.

The model's hyperparameters were tuned until the highest out-of-sample predictive accuracy achieves. To ensure that the high accuracy achieved was not specific to a training sample, I bootstrapped the training sample, re-trained the model with fresh sets of training samples, and tested its accuracy in a new validation set. This re-training-and-validation exercise were repeated three times. On average, my model achieved an *F*-score of 82%[11], suggesting that its accuracy was comparable to, and in some cases surpassing, prior studies using StockTwits data (see, e.g., Giannini, 2017; Cookson and Niessner, 2020).

3.5 Technical details on the machine learning algorithm

As the overview below is by no means exhaustive, interested readers are encouraged to attend the online course CS224N provided by Stanford University. The underlying classification

[10] I acknowledge that neural network is learning the statistical association between words and hence is a black-box in nature. However, choosing training sample to learn with correct context surpass the classification accuracy as compared to prior dictionary type measure.
[11] Using in-sample, 89% predictive accuracy is achieved.

algorithm, FastText, is a heavily modified version of Word2vec (Mikolov et al. 2013). A unique feature of Word2vec (or FastText) is that the representation of words is in the form of vectors. This feature is considered one of the most innovative ideas in the Natural Language Processing (NLP) industry. For example, the word "apple" is represented in the following vector:

$$Apple = \begin{bmatrix} 1 \\ 5 \end{bmatrix}^T$$

Where superscript *T* denotes transpose, the vector has two dimensions and thus two values (*x*=1 and *y*=5) that capture the characteristics of the word "apple", indicating its "address" in a 2-dimensional space. In practice, vectors for some words could have up to 300 or more dimensions. Now, consider a tweet written as "Apple missed consensus earnings forecast."

Figure 2
A bearish tweet

$$\underbrace{Apple}_{context\ word\ W_{t-1}} \quad \underbrace{\boldsymbol{missed}}_{\boldsymbol{center\ word\ W_t}} \quad \underbrace{consensus}_{context\ word\ W_{t+1}} \quad \underbrace{earnings}_{W_{t+2}} \quad \underbrace{forecast}_{W_{t+3}}$$

To learn the meaning of a tweet, the best way is to analyse the words that surround the centre word and the conditional probability of the centre word W_t given its neighbors W_{t-1}, W_{t+1} ($P(W_t|W_{t-1})$ and $P(W_t|W_{t+1})$). In other words, as Figure 2 shows, the algorithm predicts the likelihood of having a centre word, "missed", conditional on the presence of neighbouring words "Apple" and "consensus", an approach that is referred to as the Continuous Bag of Words (CBOW). Such conditional probability is estimated using a probability function such as Softmax.

Suppose we want to calculate $P(W_t|W_{t+1})$, that is, the probability of "missed" given the presence of "consensus" next to it, using Softmax function:

$$P(W_t|W_{t+1}) = \frac{\exp\left(\overrightarrow{W_t}^T \overrightarrow{W}_{t+1}\right)}{\exp\left(\overrightarrow{W_t}^T \overrightarrow{W}_{t-1}\right)+\exp\left(\overrightarrow{W_t}^T \overrightarrow{W}_t\right)+\exp\left(\overrightarrow{W_t}^T \overrightarrow{W}_{t+1}\right)+\exp\left(\overrightarrow{W_t}^T \overrightarrow{W}_{t+2}\right)+\exp\left(\overrightarrow{W_t}^T \overrightarrow{W}_{t+3}\right)},$$

In the numerator, $\overrightarrow{W}_t$ and $\overrightarrow{W}_{t+1}$ are the vectors of the center word "missed" and the context word "consensus". Multiplying two vectors $\overrightarrow{W_t}^T \overrightarrow{W}_{t+1}$ gives a scalar output. The larger the scalar, the more "similar" both words are, and the higher the conditional probability. In the denominator, we sum the "similarity" of all possible combinations of words in the tweet to create a probability

distribution, $P(W_t|W_{t+1})$.[12] Softmax applies exponentials to magnify (minimize) the importance of the largest (lowest) dot product similarity.[13]

Once the conditional probabilities for the centre word "missed" are computed, the centre word will be shifted to the next word, which is "consensus" in this case. Using similar procedures, the conditional probabilities for "consensus" in relation to the new neighbouring words will then be calculated. The centre word will then be shifted to the next word again and the process of estimating conditional probabilities will continue.

Figure 3
A bearish tweet

$$\underbrace{Apple}_{W_{t-2}} \quad \underbrace{missed}_{context\ word\ W_{t-1}} \quad \underbrace{\boldsymbol{consensus}}_{\boldsymbol{center\ word\ W_t}} \quad \underbrace{earnings}_{context\ word\ W_{t+1}} \quad \underbrace{forecast}_{W_{t+3}}$$

The algorithm loops through every word in the tweet until the very last word, and the process is repeated for the next tweet. For each iteration within the tweet, the goal is to minimize the sum of probability value using a negative log likelihood function (cost function). All of the above process can be summarized by this negative log likelihood function, $J(\theta)$:

$$J(\theta) = -\frac{1}{T}\log\ L(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\le j\le m}\log P\left(W_t|W_{t+j};\theta\right)\ where\ j\ \neq 0$$

Variable	Definition
t	The current position of centre word within a tweet.
T	The final position of centre word within a tweet.
θ	The word features or the vector.
m	The context window size, in our example $m = 1$.
j	The current context word, for example $j = 1$ ($j = -1$) are the context words leading (trailing) the center word.
$P\left(W_t\|W_{t+j};\theta\right)$	The conditional probability calculated by e.g. Softmax function.

Put simply, suppose we are currently at centre word t. We sum all the log conditional probability $\log P\left(W_t|W_{t+j};\theta\right)$ with the context word j surrounding t with context window size m. We sum again the moving window from center word t to T and average by number of words in the tweet $\frac{1}{T}$ to get a score $J(\theta)$ of the word vector.

[12] Since our example is a single tweet with five words, we sum all five combinations. In practice, the denominator is not restricted to a single tweet but includes the vocabulary from the full sample of tweets.
[13] Note that the vector for each word is initially generated from random values, but optimization pushes their values and word characteristics to the correct "address".

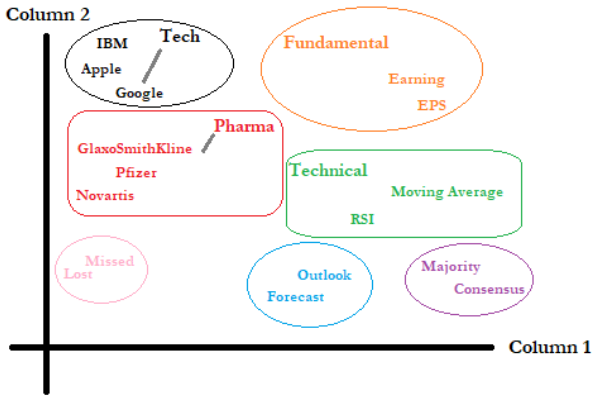
The objective of minimizing negative log likelihood function is to find the closest distance from the probability to the true (but unknown) probability. The answer to the minimization problem is to take derivatives of the cost function $J(\theta)$ with respect to word vector θ and use algorithms such as stochastic gradient descent - that is, by trial and error - until a local minimum on the gradient is reached. By performing such minimization procedures tweet by tweet, each word's vector will be updated and obtain unique features (characteristics). Such features will be stored into a matrix for each word, with rows representing the word and columns as the features.

When the whole process is finished, words will have distinct characteristics and the matrix (corpus) would appear as below:

$$\begin{matrix} Apple \\ missed \\ consensus \\ earning \\ forecast \end{matrix} \begin{bmatrix} 1 & 5 \\ 1 & 2 \\ 6 & 1 \\ 4.5 & 4 \\ 3 & 1 \end{bmatrix}$$

Graphically speaking, when the model is trained with many tweets:

Figure 4
Example of clusters



Words with similar meanings or characteristics would be clustered together (as circled in Figure 4). Their Cosine similarity[14] would also be closer. An important feature of such clustering and vector storage is word analogies; that is, our model allows us to guess or predict the word associated with another by adding or subtracting vectors. For example, if "Technology" is associated with "Google," and we ask what "Glaxo Smith Kline" is associated with, a well-trained model likely returns "Pharmaceutical."

[14] Cosine similarity measures the angle between two vectors.

Understanding the meaning of multiword phrases requires the modelling of conditional dependence. For instance, that "not good" is analogous to "bad" could be learned by expanding the context window to bi-gram, tri-gram to *n*-gram; that is, by changing the context-size hyperparameter m. Furthermore, tweets may have irregular characteristics, as users may use abbreviations such as "w/o" or "w/" to denote "without", "b/c" for "because", etc. Furthermore, users may use long or extended words to express intensified emotions, such as "Goooood" to make an emphatic expression that is beyond "good." FastText recognizes the meaning by analysing sub-words, that is, performing the previous word-vector minimization procedure at the character level, which allows the algorithm to learn and relate such intensified or abbreviated words to the original words.

Once FastText has obtained a large corpus of optimized word-vector characteristics and similarities, the next step is then to classify tweets into three labels - bullish, neutral, and bearish. For FastText to learn and classify these tweets into these labels, pre-classified tweets - those that the users have explicitly labelled as bullish, neutral, or bearish - are used to train the data in-sample. Using the large word-vector matrix obtained previously, all word-vectors within tweets are first averaged to the tweet-vector level. A multinomial logistic regression that regresses the three labels on the tweet-vector characteristics is estimated for the training data. Using the estimated parameters, I obtained the predicted probabilities for the unlabelled tweets and classified them into three labels.

80% as a training set for these pre-classified tweets were held off and the other 20% as a validation set. I also made the training set balanced for each label because the logistic model would often favour the prediction outcome of the label with a significantly higher number of observations, as it had more information to learn. Moreover, it reduced the likelihood of having type I or II errors when unlabelled tweets were classified into labels.

Using the training set, I fine-tuned the FastText hyperparameters, including the learning rate, number of epochs, context window size, and sub-word information, until the best predictive accuracy was achieved.[15] Once optimal accuracy was achieved, its robustness was confirmed by retraining the model and bootstrapping pre-classified tweets three times to ensure that each training and validation sets had some new tweets.

[15] Our choices of hyperparameters used in this project are available upon request. More information on parameter tuning can be found in fastText documentation: https://fasttext.cc/docs/en/support.html

3.6 Modelling approach

StockTwits is a panel data; it contains both cross-section and time-series elements. There are two primary models used for estimation, namely, the fixed-effects model and the random-effects model. Assumed that the dataset is two-dimensional, namely user and time level, the user and time fixed components exist wherein user level, and the differences in data occur across user dimensions but not across time. On the other hand, along with time level, data differs along time dimension but not at user dimension. Often, researchers who choose to use fixed effect modelling will control for both user and time, for example, year-month, fixed-effect.

3.6.1 Fixed effect model

To illustrate the idea of fixed-effect, consider a user fixed-effect model given by this equation:

$$y_{it} = \beta_0 + \beta_1 x_{it} + user_i + \epsilon_{it}$$

Where ϵ_{it} is the error term with zero mean and constant variance, and $i = 1, \dots, N$, $t = 1, \dots, T$ represents user and time, respectively. While x_{it} is observed, $user_i$ is an unobserved time-invariant variable.

An important characteristic of the user fixed-effect model is that it allows $user_i$ to be correlated with other independent variables x_{it}:

$$Cov(user_i, x_{it}) \neq 0$$

Note that strict exogeneity with the error term ϵ_{it} is still required.

Simply put, user fixed-effect model allows a user's time-invariant innate ability, such as race, gender, IQ, etc., to be correlated with other independent variables. For example, in the application of stock return prediction, a user's time-invariant characteristics like IQ can affect the $Sentiment$ and are correlated with each other.

Since $user_i$ is not observable, it cannot be directly controlled for. The fixed-effect model eliminates the time-invariant term $user_i$ with two main estimation methods. The first method is by including a dummy variable for each user $i > 1$. This is numerically equivalent to the fixed-effect model and only applicable when degrees of freedom are not exhausted after including the $user_i$ dummies. In other words, adding dummy variables will control for differences across users such that each user will have a slope and intercept in the model..

The second method is called within transformation or demeaning the data, that is, by averaging the data across time and calculating the difference to the mean:

$$y_{it} - \overline{y_i} = \beta_0 - \overline{\beta_0} + \beta_1(x_{it} - \overline{x_i}) + (user_i - \overline{user_i}) + \epsilon_{it} - \overline{\epsilon_i}$$

Where $\overline{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\overline{x_i} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$, and $\overline{\epsilon_i} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}$

By demeaning $y_{it}, x_{it}, \epsilon_{it}$ the model is transformed by eliminating the intercept $\beta_0 - \overline{\beta_0}$ and user fixed effect $(user_i - \overline{user_i})$ since β_0 and $user_i$ are constant. The model in both methods can be estimated by the usual OLS model and the parameter of interest β_1 are the same as the underlying model.

3.6.2 Random effects model and between estimators

Conversely, the random-effects model in panel data assumes that the differences between users are assigned randomly instead of fixed. This is modelled by including a fixed intercept $\overline{\beta_0}$ and a random variable $user_i$ which varies across individuals. This $user_i$ is assumed to have zero mean and constant variance and is the same as the error term ϵ_{it}:

$$y_{it} = \overline{\beta_0} + \beta_1 x_{it} + user_i + \epsilon_{it}$$

At first glance, the model appears similar to the fixed-effect model. However, the main difference is that under the random-effects model, the $user_i$ the unobserved user time-invariant component cannot correlate with other independent variables x_{it}:

$$Cov(user_i, x_{it}) = 0$$

If this assumption holds, the random-effects estimator is more efficient than the fixed-effects model because it is estimated with GLS than OLS. The variance will generally be smaller. However, if this assumption does not hold, the random effects estimator is not consistent.

On the other hand, if researchers are interested in the time-invariant variable, but fixed-effect (within estimator) is not applicable since it removes such variable, there is another type of model called Between estimator. Between estimators eliminate the time dimension by averaging, and identification relies solely upon the cross-section, for example:

$$\overline{y_i} = \beta_1 \bar{x}_i + user_i + \overline{\epsilon_i}$$

Where $\overline{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\overline{x_i} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$, $\overline{\epsilon_i} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}$. Note that by averaging the time dimension, all variation between time will be lost.

The problem arises in Between estimators because $user_i$ the unobserved user time-invariant components still exist and must satisfy the zero correlation with other independent variables as a consistent estimator. Even if one meets the restriction pointed above, the Between estimator is useful in considering the random effects instead as an estimator in its own right.

3.6.2 Fixed effect as the predominant model in finance and economics literature

To the best of my knowledge, recent finance and economics literatures have hardly any application that satisfies $Cov(user_i, x_{it}) = 0$. For instance, as discussed in the fixed-effects model, a user's time-invariant ability, like IQ, will be correlated with $Sentiment$ variable. In corporate finance, a firm's corporate culture inherently affects a firm's performance like firm size, and profitability. In economics, a country's or state culture and history can affect a country's foreign direct investment (FDI), etc. StockTwits is a panel data with three dimensions: user, firm, and time and hence, fixed-effects modelling is used for the analysis in the following chapters.

To illustrate the idea that most studies use fixed-effect modelling, the table below shows a non-exhaustive list of literature and their empirical methods:

Literature	**Main result identification**
Limited Attention	
Corwin and Coughenour (2008)	Firm fixed-effect
Sonney (2009)	Firm-year fixed-effect
Dellavigna and Pollet (2009)	Time fixed-effect
Da et al. (2011)	Time fixed-effect
Baker and Wrugler (2006)	Pooled OLS
Custodio et al. (2013)	Firm-year fixed-effect
Custodio et al. (2019)	Firm-CEO fixed-effect
Jung et al. (2012)	Industry and year fixed-effect
Harford et al. (2018)	Analyst, firm, or analyst-firm fixed-effect
Dunn and Mayhew (2004)	Pooled OLS
Disagreement and short-selling constraint	
Almazan, Brown, Carlson, and Chapman (2004)	Pooled OLS, Random-effect, Between estimator
Diether, Malloy, and Scherbina (2002)	Portfolio Sorting
Dischse (2002)	Portfolio Sorting
Serafeim and Yoon (2021)	Industry and time fixed-effect
Mankiw, Reis, and Wolfers (2004)	Pooled OLS
Lamont (2002)	Individual forecaster fixed-effect
Frazzini and Pedersen (2014)	Portfolio Sorting
De Giorgi, Post, and Yalcin (2020)	Portfolio Sorting
Barinov (2014)	Portfolio Sorting
Social-Media Sentiment	
Antweiler and Frank (2004)	Firm fixed-effect
Bartov, Faurel, and Mohanram (2018)	No mention of identification
Cookson and Niessner (2020)	Time fixed-effect
Giannini, Irvine, and Shu (2017)	Firm fixed-effect
Ivkovic, Zoran, and Weisbenner (2005)	Area fixed effect
Tetlock (2007)	Pooled OLS

Almazan, Brown, Carlson, and Chapman (2004) is a paper about mutual fund managers constrained by short-selling and is the only paper in the list that use all Pooled OLS, Random-effect, and Between estimator. To understand why there is a problem with the unconventional methods, they examine what causes the fund manager to constrain themselves in investment practices and proposed a monitoring hypothesis where the fund management team has a moderating effect on the fund manager's action. The data is from SEC form N-SAR question 70, where the fund manager is required to answer "Yes" or "No" on six investment-specific questions. The categories are: "borrow of money", "margin purchase", "short-selling", "writing or investing in options on equities", "writing or investing in stock index futures", and "investments in restricted securities". The dependent variable is then a Score from equal-weighting the six categories in total to reveal how constrained a fund manager is.

The main result regression is presented below:

$$\begin{aligned} Score_{it} = {} & \beta_0 + \beta_1 BoardSize_{it} + \beta_2 PropIndep_{it} + \beta_3 Team_{it} + \beta_4 MgrAge_{it} \\ & + \beta_5 FrontLoad_{it} + \beta_6 BackLoad_{it} + \beta_7 Top10_{it} + \beta_8 LogFundAge_{it} \\ & + \beta_9 LogTNA_{it} + \beta_{10} Turnover_{it} + \beta_{11} LB_{it} + \beta_{12} LG_{it} + \beta_{13} MV_{it} + \beta_{14} MG_{it} \\ & + \beta_{15} SV_{it} + \beta_{16} SB_{it} + \beta_{17} SG_{it} + \epsilon_{it} \end{aligned}$$

Where i and t represent fund and year, respectively. $\{BoardSize_{it}, PropIndep_{it}, Team_{it}, MgrAge_{it}, FrontLoad_{it}, BackLoad_{it}, Top10_{it}\}$ are the managerial hypothesis variables that relate to the fund management, for example, management board size ($BoardSize$), manager's age ($MgrAge$), the intensity of direct monitoring by the board ($PropIndep$), etc. The rest of the control variables are the fund performance such as the log of Total Net Asset ($logTNA$), fund's turnover $Turnover$, log of fund age ($logFundAge$), etc.

In terms of estimation, there are four alternative regressions, all with fund-year observation. To justify the use of Pooled OLS, they suggest "… Models run by pooling fund-year observations can produce more efficient estimators because they take advantage of the panel structure of the entire sample." Using Pooled OLS by itself is misspecified because a time-invariant fund characteristic, for instance, investment style and the manager himself subsumed in fund fixed-effect, will correlate with other fund performance and the fund manager's characteristics. The use of other Random-effect and Between estimators will also be inconsistent. Recently, fund studies such as Pastor, Stambaugh, and Taylor (2017) and Wang (2019) have proposed to use the fixed-effect model.

3.7 Summary for Chapter 3

Having introduced the data and method used in this thesis, I examined the limited attention and disagreement as the goal reviewed in chapter 2, which are the behavioural biases faced by individual investors.

The primary data used is from a social media company called "StockTwits". "StockTwits" is a social media platform that focuses only on the financial market and has been the primary place for individual investors to exchange trading ideas, opinions about stocks. I downloaded 48 million tweets posted by more than 280,000 unique registered users from January 2010 to December 2017. In each tweet, I collected the text content, company tickers, and the ID of users who tweeted. Any information unrelated to the stock and tweet without referring to a valid company ticker was removed. User-level data such as users following, or being followed by, user number of likes received from other users were also downloaded.

Next, the concern that some users were, in fact, "social-media bots" run by computer algorithms had also been addressed, because these bots would indeed post informative tweets regarding stocks. However, since the goal is to examine individual investors, these bot users were removed by detecting abnormal user characteristics proposed by (Efthimion, 2018; Kearney, 2018) which were suggestive of bot users. Overall, 6,354 users were removed, reducing the number of users to 277,389. Furthermore, I followed prior literature and performed data cleaning steps for the tweets by removing tweets containing conversational dialogues, website URLs, or tweets written in non-English characters. These tweet samples were then fed into a machine-learning algorithm called "FastText" to measure the bullishness or bearishness.

To evaluate the best approach to classify the tweets, I discussed the pros and cons of various machine learning and dictionary-based tweet classification methods. I provided the reasons, steps of using "FastText" as the choice of the machine learning algorithm. The technical details of "FastText" inner workings were also explained to increase transparency to readers unfamiliar with machine learning literature.

Lastly, I discussed the modelling approach using StockTwits. StockTwits which was inherently a panel data with three dimensions: user, firm, and month. I also discussed the advantage and disadvantages of using fixed-effect and random-effects modelling and how fixed-effect was a more appropriate approach to model the dataset.

4. Limited Attention in Stock Returns: Evidence from Social Media

4.1 Introduction

This chapter focuses on individual investors and examines the implications of generalization and specialization approaches for the quality of their stock analysis. Individuals who prefer following and analysing a wide range of stocks or stocks from various industries are defined as having a *generalist* approach. In contrast, those who follow and study only a few stocks have a *specialization* approach.

The longstanding *generalist-specialist* debate has been examined in a variety of contexts. For example, Economists study the value implications of such taxonomy by analysing communication quality between individuals with different skill sets (Ferreira and Sah, 2013). In the finance literature, some have examined managerial skills developed in prior work experience and reported that generalist managers receive higher pay (Custodio et al., 2013) and achieve greater innovation success (Custodio et al., 2019). Sonney (2009) examined this debate in the context of analyst research and showed that earnings forecast is more accurate for analysts with a country specialization than those with industry specialization.

Research consumes considerable time and effort, but those who spend resources gathering information are compensated (Grossman and Stiglitz, 1980). In the context of equity investment, researching stocks typically involves gathering company information, analysing and interpreting qualitative and numerical data, and exercising subjective judgment in formulating views and predictions about fundamental values and future prices (Jung et al., 2012; Harford et al., 2018).

Due to the complex nature of such processes, the effect of generalization or specialization on individual investors' efficacy of stock analysis is ambiguous. Conversely, individuals with a generalization approach likely possess superior industry knowledge- and market-wide factors and thus have an informational advantage over those who are relatively more specialized (Dunn and Mayhew, 2004; Sonney, 2009). On the other hand, since human beings have limited cognitive resources to apply to different tasks, the attention spent on one task necessarily reduces attention available for other tasks (Kahneman, 1973). As such, one cannot maintain the same level of time and effort on research as the number of stocks followed/analysed increases. Hence, compared to specialists, equity research by generalists may be inferior because the time and effort allocated to each stock on their “watch” list is likely to be reduced, implying that company

information may not be fully gathered, processed, and synthesized (Boni and Womack, 2006; Bradley et al., 2016). By the same token, investors following only a few stocks likely pay more attention to each and thus have more in-depth and potentially superior knowledge (Clement, 1999; Gilson et al., 2001; Corwin and Coughenour, 2008).

Alternatively, it could be argued that specialists are more likely to be subject to behavioural bias, such as developing myopic and overly-optimistic views of companies they have spent considerable time and attention on, and with which familiarity and personal attachment may have developed (Piotroski and Roulstone, 2004; Chan and Hameed, 2006). The question of whether a generalization or specialization strategy in equity research fares better for individual investors is ultimately an empirical question.

It would seem logical that the ideal database to test this question is one containing actual equity trading information by individual investors. However, such data is not readily available. In addition, since only trades are recorded in such databases, the investors' stocks that are considered, followed, and analysed but not eventually traded do not enter the data. Hence, their orientations in stock coverage (generalization vs. specialization) cannot be reliably measured.[16] Additionally, gauging the number of stocks in an equity portfolio would predominately capture the user's preferences in diversification rather than the approach to following and studying stocks, although both would be positively correlated. To address these challenges, I proposed a relatively cleaner empirical setting, StockTwits, one of the largest social-media platforms in security trading, to test the research question.

StockTwits is a social-media platform on which users post short, micro-blog messages (tweets) about companies in order to share trading ideas and insights with other users. Since its establishment in 2010, StockTwits has gained popularity, secured a large user base, and has been employed by professional financial-service firms such as Bloomberg, to analyse market sentiment data and other news feeds. I have downloaded all tweets on StockTwits over the period 2010 to 2017 and believe that such data is instrumental to testing our research questions for at least three reasons.

[16] Nonetheless, one could reasonably point out that there are cases where users do not tweet all stocks that they follow, study, or trade. As such, the set of companies users tweet about may not fully capture the breadth of their stock coverage. While this concern is valid, we believe that this is unlikely to affect our conclusions because stocks tweeted about are likely to be ones that users are paying active attention to and showing greater interest in. Since the later part of our analysis focuses on the role of investors' attention, the set of tweeted stocks, though measured with errors, gives a reasonably good estimate of the full set of stocks users are paying attention to and following actively.

First, since users often post tweets about stocks they are interested in and have been watching, data from StockTwits enables me to trace the set of stocks individual users follow, pay active attention to, and analyse over time. Second, by estimating social-media sentiment using Natural Language Processing techniques and studying its predictability over future stock returns, this setting benefits from having a relatively objective benchmark for performance when comparing users with a generalization approach to those who prefer specialization. Third, the user-level data allows me to include user-fixed effects in the estimation, thereby differencing out between-user variations, such as talent, trading styles, financial sophistication, etc., which are likely confounding factors OLS estimation.

Using the StockTwits sample consisting of more than 156,000 users, covering 8,575 U.S. publicly-traded firms over the period of 2010-2017, I find that social-media sentiment predicts positively and significantly future monthly stock returns, controlling for a wide array of firm characteristics, and user and month fixed effects, which complements prior studies (see, e.g., Renault, 2017; Bartov et al., 2018). Importantly, the tests reveal that such positive predictability decreases significantly with the number of stocks users have been following over the previous six months, consistent with a limited-attention behavioural bias explanation. Economically, the estimates show that when the number of stocks covered by users is at the 25^{th} and 75^{th} percentiles, a one-standard deviation increase in social-media sentiment is associated with increases in future stock returns of 37.8 and 17.0 basis points per month, respectively. These results are robust to the use of risk-adjusted stock returns and alternative windows for measuring stock and industry coverage, industry classifications, weighting schemes for estimating social-media sentiment, and sample exclusion restrictions. Further tests confirm that the effect in question is independent from that of user reputation and of prior experience with the followed stocks.

I also performed three tests to examine the role of limited attention in driving our results. First, the limited-attention theory asserts that users spend time and attention gathering and analysing fundamental information about their followed companies. Since certain types of users such as those trading primarily on technical analysis, as well as those who have very short-term investment horizons tend to place little emphasis on a company's fundamental information in formulating trading ideas or views, a behavioural explanation thus requires that those traders do

not drive our estimation results. Excluding users who self-declare themselves to be professional, technical, or day traders does not affect our baseline results.

Second, to the extent that users following many stocks allocate less time and attention to analysing the fundamental positions of their followed companies, social-media tweets posted by such users are less likely to be reliable when predicting future earnings performance, especially when compared to those posted by specialist users. In line with this conjecture, when the number of stocks covered by users over the past six months is at the 25^{th} and 75^{th} percentiles, a one-standard-deviation increase in social-media sentiment is associated with \$0.008 and \$0.001 higher earnings per share, respectively. These results continue to hold when earnings surprises are examined.

Third, I explored whether the predictive power of social-media sentiment over future stock returns depended on company characteristics that clearly capture how users comprehended their financial positions. If limited attention drives our findings, the negative moderating effect of stock coverage is likely to be more pronounced among informationally-opaque companies with complex organizational structures, which are harder to analyse. Among such stocks, the loss in accuracy for stock-return predictions is likely to be more severe when research effort and effectiveness is reduced due to users' attention and time constraints. Results from subsample tests supported this conjecture. Overall, the empirical results are consistent with the limited-attention hypothesis.

Finally, a trading strategy was designed to exploit the return predictability of social-media sentiment. Users were divided into low and high stock-coverage groups; the low group consisted of users covering 100 stocks or below, and the high group contained more than 100 stocks. I equal-averaged social-media sentiment for each of the two user groups and computed the net (low-minus-high) sentiment for each stock. A trading strategy that longs a portfolio of stocks with zero and positive net sentiment, and shorts a portfolio of stocks with the negative net sentiment, generates a significant monthly excess return of 21.7 basis points and a trading alpha (based on Carhart (1997) 4-factor model) of 26.4 basis points per month.

This study contributes to the literature in several ways. Firstly, the findings added to the growing body of literature examining the implications of investor sentiment for asset prices (see, e.g., Baker and Wrugler, 2006; Baker et al., 2012; Da et al., 2015). Numerous studies now estimate social-media sentiment at the firm level. For instance, measuring sentiment using

Twitter data, Bartov et al. (2018) examined its link with future earnings and stock returns. Cookson and Niessner (2019) investigated how disagreement between StockTwits users drives trading volume. Giannini et al. (2017) demonstrated how the social-media sentiment of StockTwits users who are geographically close to the companies has stronger predictive power over future stock returns than those further away, consistent with the former possessing advantageous local information. My findings complement the above studies. Importantly, this study also uncovers a new user characteristic - the number of stocks followed - that determines the predictive power of sentiment over future stock returns. A trading strategy exploiting the variation in such a user characteristic generates significant trading profits.

Secondly, I further contributed to the behavioural finance literature by examining how attention constraints affect economic decision-making. For instance, Peng and Xiong (2006) showed that cognitively overloaded investors exhibit "category-learning" behaviours tending to process more market- and sector-wide information, which could explain asset-price dynamics not covered in rational expectations models. Hirshleifer et al. (2011) presented a model in which inattentive investors receive only partial firm-specific information and make losses in the equilibrium. On the empirical side, using individual NYSE specialists' portfolio size to proxy for attention span, Corwin and Coughenour (2008) found that liquidity provision decreases as portfolio size increases, consistent with a limited-attention explanation. My findings present new evidence that limited attention plays a significant role in determining the quality of stock analysis by individual investors, especially when the companies they follow have complex organizational structures and are informationally opaque.

Third, I offered new insights into the longstanding debate on whether a generalization or a specialization approach fares better (see, e.g., Gilson et al., 2001; Sonney, 2009; Custódio et al., 2013; Custódio et al., 2017). In the context of stock investment, much of the discussions along with this line focus on diversification: the question of how many stocks an investor should put in their portfolios to achieve diversification (see, e.g., Statman, 1987; Statman, 2004). However, relatively few studies focus on stock analysis - an important process before an investor formulates their investment decisions - and several questions remain open. For instance, relatively little is known about how many stocks individual investors prefer to "watch" and follow concurrently, and more importantly, whether such preference may determine their investment performance. Hence, in the presence of behavioural constraints, the more specialized

an individual user is, the better the quality of their equity analysis. I shed light on these questions by demonstrating that social-media tweets predict future stock returns less accurately when the number of stocks users follow increases.

Fourth, I contributed and served confirmation to the experimental finance literature, particularly in Bossaerts et al. (2020). Their paper studies individual computational complexity where assets uncertainty is the consequence of not knowing which information is pertinent rather than not having the information in the first place. They formalize this as a 0-1 Knapsack problem, a combinatorial optimization problem whereby asking the decision-maker to find the subset of items of different values and weights that maximize the total value knapsack subject to a weight constraint. In the laboratory experiment, individuals endowed with several securities trade to maximize this knapsack problem. The result is consistent with my finding where price information generally revealed incorrection solutions, and the informational quality deteriorated as the instance complexity increased.

Lastly, I contributed to the literature by incorporating asset pricing, behavioural finance, and corporate finance. My econometric method is more inclined towards the latter twos and is different from standard asset pricing measurement, where they look at pricing factors. In the baseline analysis, apart from looking at Social Media sentiment and other firm controls as pricing factors, I also included a behavioural component of user ability to predict return with his/her stocks coverage. In fact, the relationship between systematic risk and the expected return of asset pricing are based on efficient, competitive market and risk adverse investers which have been shown not to hold in reality; hence the study of behavioural finance in terms of limited attention and disagreement and corporate finance with the fixed-effect regression that focuses on *individual investor* instead of *firm* behaviour might bring such relationship closer to reality. The subsequent analysis also seeks to understand the channels where the informational opacity from the firm presents, leading to an increase in computational complexity of the investors in general.

4.2 Measuring StockTwits user-level social-media sentiment and their coverage

Measuring Social-media Sentiment

This section discusses how user-level social media sentiment is measured. StockTwits' tweets are measured and classified into {Bullish, Neutral, Bearish} sentiment as outlined in section 3.3. "Neutral" tweets in which the user does not talk about the firm are deleted as they introduce "noise" to the *Sentiment* variable. The rest of Bullish and Bearish tweets are aggregated into the *Sentiment* posted by each user in a month.[17] The overall sentiment of firm *j* given by user *i* in month *t* is calculated as the number of bullish tweets posted divided by total bullish and bearish tweets added together, as follows:

$$Sentiment_{i,j,t} = \frac{\#\ of\ bullish\ tweets_{i,j,t}}{\#\ of\ bullish\ tweets_{i,j,t} + \#\ of\ bearish\ tweets_{i,j,t}} \quad (1)$$

The sentiment measure, $Sentiment_{I,j,t}$ is bounded between 0 and 1, with values above (below) 0.5, indicating a user's overall bullish (bearish) sentiment in a given month.

This sentiment measure is adopted from Antweiler and Frank (2004), where they defined as $Sentiment_{j,t} = \frac{\#\ of\ bullish\ tweets_{j,t} - \#\ of\ bearish\ tweets_{j,t}}{\#\ of\ bullish\ tweets_{j,t} + \#\ of\ bearish\ tweets_{j,t}}$. Their approach differs from equation (1) for two reasons. First, since they are not focusing on limited attention but simply the predictivity of *Sentiment* on future returns, they aggregate *Sentiment* to a firm-month level. My measure aggregates to a user-firm-month level to study the role of limited attention on the user level. Secondly, their approach deducts # of bearish tweets in the nominator of *Sentiment* compared to equation (1) and essentially rescales from -1 and +1. I adapted their approach to confirm that equation 1 is robust, and the regression results are *exactly* the same.

Antweiler and Frank (2004) also proposed different ways of aggregating sentiment, for example:

$$Sentiment_{j,t} = \ln\left[\frac{1 + \#\ of\ bullish\ tweets_{j,t}}{1 + \#\ of\ bearish\ tweets_{j,t}}\right]$$

$$Sentiment_{j,t} = \#\ of\ bullish\ tweets_{j,t} - \#\ of\ bearish\ tweets_{j,t}$$

[17] For tweets posted during non-trading days and after market close, I followed Cookson and Niessner (2020), whereby user *i*'s day *t* tweets are defined as end-to-end market close (4 pm) from day *k-1* to day *k* (4 pm). If tweets are in non-trading days, e.g. public holidays, and weekends, the date of these tweets will be rolled to the next available trading day. If there is no tweet in a given user-firm-day, that is, no tweet in both "Bullish" and "Bearish" labels, we omit the whole day of data point of user-firm-day level data.

The latter aggregation method is used by studies like Giannini et al. (2017) and Bartov et al. (2018). Both measures take into account the number of tweets expressing a particular sentiment. In other words, the more tweets, the higher the *Sentiment* value. I also followed their alternate aggregation methods in an unreported table and found consistent results in equation (1).

Measuring stock and industry coverage of StockTwits users

To capture the breadth in stock and industry coverage of StockTwits users, I traced each user over time and counted the number of unique stocks they have tweeted about (*# of stocks covered*). Since the goal is to identify the set of stocks and industries that users are interested in and pay attention to simultaneously, I chose 6 months for the time period and present sensitivity tests based on a 3-month window in the robustness section.[19]

The second measure of coverage is the number of unique 3-digit SIC industries, spanned by the set of unique stocks followed by users over the 6-month period (*# of industries covered*). In subsequent robustness analysis, I applied alternative industry classifications, such as the Fama-French 49 industries and 4-digit SIC industries, in constructing this measure and found similar results.

[19] If the time period is set too long, the likelihood of users changing interests and the set of stocks currently followed would be high. As such, we have a higher risk of counting stocks that are not followed by users anymore. Likewise, if the period is set too short, we run the risk of omitting stocks that are followed by users but not tweeted about in the recent past.

4.3 Summary and Descriptive statistics

Table 3 panel A shows the descriptive statistics for the StockTwits sample from Jan 2010 to Dec 2017. There are 29,991,466 tweets used in this analysis, of which 16,783,135 tweets are classified as Bullish, and 13,208,331 are *Bearish*. In total, 156,675 unique active users are covering 8,575 unique firms. On average, a user posts 4.956 tweets per month, of which 2.738 (2.218) are *Bullish* (*Bearish*) tweets.

TABLE 3
Descriptive Statistics

This table presents descriptive statistics for our sample of tweets from StockTwits. The sample period is from January 2010 to December 2017. Panel A gives the number of tweets by message type and for an average user, the number of stocks covered by all users, and further statistics on when maximums and minimums per month. Panels B and C show the average (user-average of time-series per month) user number of tweets and unique users by the number of stocks and the number of industries covered.

Panel A. Descriptive statistics of stock tweets

	Total	Average of tweets per month	Stdev.	25%	Median	75%
# of tweets	29,991,466	333,239	210,429	161,448	306,809	454,318
# of Bull. tweets	16,783,135	186,479	83,207	113,393	207,268	257,786
# of Bear. tweets	13,208,331	146,759	152,007	42,043	73,501	195,070
# of unique and eligible users	156,657					
# of firms covered by all users	8,575					

	User-average of time-series means
# of tweets an average user post	4.956
# of bull. tweets an average user post	2.738
# of bear. tweets an average user post	2.218

	Values	Months
Highest # of active users	37,085	2017m11
Lowest # of active users	1,558	2010m08
Most bullish month	0.78	2014m02
Most bearish month	0.35	2017m08
Largest # of stocks covered by all users	5,301	2016m07
Smallest # of stocks covered by all users	2,201	2010m07

Panel B. Statistics by stock coverage

	# of stocks covered					
Average of time-series per month	1-20	21-40	41-60	61-80	81-100	100+
# of tweets posted	4.85	4.59	4.22	3.78	3.69	3.01
Sentiment	0.65	0.63	0.63	0.64	0.64	0.68
Bullish tweets	3.11	2.77	2.48	2.24	2.20	1.99
Bearish tweets	1.75	1.83	1.74	1.54	1.49	1.02
# of unique users	10620.21	1359.03	468.31	226.72	126.21	376.41

Panel C. Statistics by industry coverage

	# of industries covered					
User-average of time-series per month	1-20	21-40	41-60	61-80	81-100	100+
# of tweets posted	4.83	3.99	3.36	2.99	2.68	2.91
Sentiment	0.64	0.64	0.66	0.69	0.69	0.68
Bullish tweets	3.04	2.37	2.06	1.93	1.78	2.01
Bearish tweets	1.78	1.63	1.30	1.06	0.90	0.90
# of unique users	11876.87	854.11	225.29	93.87	49.24	77.52

Panel B of Table 3 divides StockTwits users into different groups (1-20, 21-40, 41-60, 61-80, 81-100, 100+) according to their *# of stocks covered*. The lowest coverage group (1-20) has the most significant number of unique users (more than 10,000 users). Such users appear to be active and have posted the greatest number of tweets on average over time compared to other groups. No apparent patterns are found in social-media sentiment across the groups. Similar results are found in Panel C, where users are divided into bins according to *# of* industries covered. Untabulated analysis reveals a pairwise correlation between the two coverage measures with a coefficient of 0.91.

Following the limited attention hypothesis, a concern was raised about Panel B where the *Sentiment* should be reducing as *# of stocks covered* increases instead of staying relatively constant. A simple example helps illustrate this idea. Suppose users A and B cover 100 and 10 stocks, respectively. They both gave *Sentiment* to two firms, "Apple" and "IBM." The *Sentiment* is a measure of bullishness and bearishness and there is no reason *Sentiment* would monotonically reduce as more stocks the user covered. The *Sentiment* may include information about the performance of the firm, macroeconomic environment, etc. When averaging the *Sentiment* of both "Apple" and "IBM" to show in Panel B, it is evident that *Sentiment* will not differ much across groups.

To see the time-series variation over time, I plotted the average tweets per user over time in Figure 5. The number of tweets a user posted has been increasing over time, as seen by a straightly monotonic increasing trend in the solid line. Users also post more bullish tweets until 2015, and bearish tweets have gained momentum since. Interestingly, in mid-2016, bearish overtake bullish tweets and coincided with the difference in opinion during the 2016 US presidential election. However, bullish tweets appear to level off in early 2017, with the highest number of tweets of 3.4. Figure 6 shows an exponential increase in users engaged in posting tweets from the lowest in early 2010 with 2,500 unique users to the highest in late 2018 with about 36,000 unique users per month. Figure 7 shows the coverage of unique firms by StockTwits users. The highest increase of users' firm coverage appeared in mid-2010 to early 2011 from 2,300 to 3,900 firms. A stable growth can be seen from 2011 to 2017, toping the highest of 5700 firms covered, consistent with the rising popularity of StockTwits over time.

Figure 5
Average Tweets Per User Over Time

Fig. 5 shows the average number of Bullish, Bearish, and their sum (Total) tweets posted by a user in my StockTwits sample from Jan 2010 to Dec 2017.

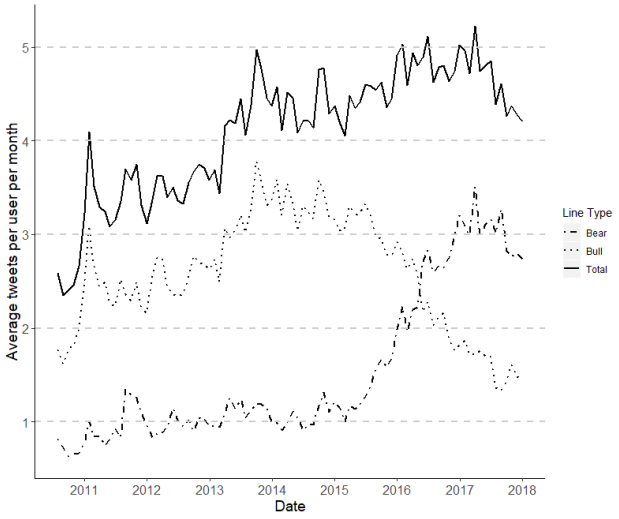


Figure 6
Number of Unique Users Over Time

Fig. 6 shows the number of unique users engaged in posting tweets in my StockTwits sample from Jan 2010 to Dec 2017.

Figure 7
Number of Unique Firms Over Time

Fig. 7 shows the monthly number of unique firms that can be successfully matched to the CRSP database in my StockTwits sample from Jan 2010 to Dec 2017.

Table 4 reports Summary Statistics of the main analysis used. Sentiment's mean (median) is the average of all observations, which is 0.66 (0.94), consistent with the user posting more bullish tweets towards the bullish side. The standard deviation of *Sentiment* is also very high with 0.44, implying users have high differences in opinion towards stocks. The mean of *# of stocks covered* is 271, whereas its median is 37, suggesting that its distribution is highly skewed. Given that nonlinear, positively skewed, and contains outliers, the natural-logarithm is used to transform the coverage measure in my regression analysis. User also tends to post on average 4.28 tweets and varies by 16 tweets. On average, users will follow a stock for about eight months, with 33 months at 95 percentile. For the firm characteristics, users follow the stocks with a market capitalization *ME* of 6.58 billion on average. The momentum factors *RET(-2, -3)*, *RET(-4, -6)*, *RET(-7, -12)* have a mean return of 2.21%, 3.22%, 7.05%, respectively, consistent with trend reversal theory where months further away from current month receives higher cumulative return than near term momentum return. *Abnormal RET* is the returns adjusted with Carhart (1997) 4 factors model shows a negative return on average of -0.08%.

TABLE 4
Summary Statistics

This table reports summary statistics for the main variables used in this study. The number of observations, means, standard deviations, and percentile statistics are reported.

Variables	Obs.	Mean	Stdev.	5%	25%	Median	75%	95%
Tweet variables (user-firm-month level)								
Sentiment	5,556,568	0.66	0.44	0.00	0.00	0.94	1.00	1.00
# of stocks covered	5,556,568	270.77	625.28	2.00	10.00	37.00	180.00	1625.00
# of industries covered	5,556,568	52.78	71.79	1.00	7.00	20.00	67.00	234.00
# of tweets	5,556,568	4.28	16.00	1.00	1.00	1.00	3.00	14.00
# of likes	5,556,568	604.46	2630.52	0.00	2.00	32.00	277.00	2469.00
# of months each stock followed	5,556,568	7.93	11.92	1.00	1.00	2.00	10.00	33.00
Firm characteristics (firm-month level)								
ME (in billion)	294,910	6.58	2.37	0.03	0.23	1.01	3.88	27.4
VOL (in thousand)	294,910	321	1206	3	27	90	273	1234
PRC	294,910	33.13	69.40	1.31	7.09	19.43	42.28	96.77
BM	294,910	2.92	67.36	0.05	0.28	0.55	0.96	3.10
Dividend Yield	294,910	0.16	0.19	0.00	0.00	0.09	0.26	0.53
RET(-2, -3) (%)	294,910	2.21	21.43	-26.78	-7.33	1.54	10.09	31.43
RET(-4, -6) (%)	294,910	3.22	26.25	-32.53	-8.88	2.29	13.03	39.34
RET(-7, -12) (%)	294,910	7.05	38.98	-43.24	-11.77	4.63	20.70	61.33
RET (%)	294,910	1.43	15.98	-19.02	-5.22	0.88	6.92	22.32
Abnormal RET (%)	226,401	-0.08	13.78	-17.95	-5.74	-0.47	4.66	18.11
Industry-adjusted RET (FF49) (%)	294,910	0.95	15.02	-17.21	-4.76	0.23	5.48	19.92
Industry-adjusted RET (SIC3) (%)	294,910	0.90	14.65	-16.60	-4.29	0.00	4.97	19.26

4.4 Social-media sentiment, Stock Coverage, and Stock Returns

4.4.1 Main analysis

To examine whether the number of stocks or industries covered by users affects the stock-return predictability of their social-media tweets, the following regression model is estimated at the user-firm-month level:

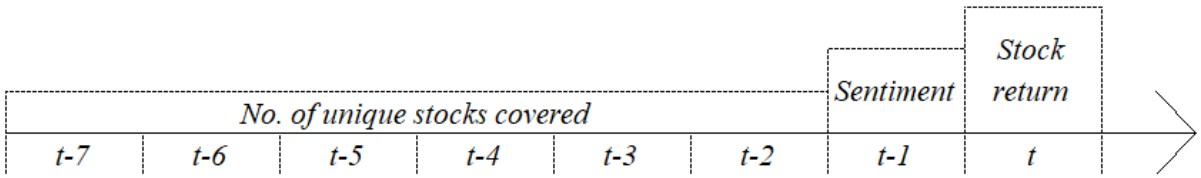
$$RET_{j,t} = \beta_0 + \beta_1 Sentiment_{i,j,t-1} + \beta_3 ln(\#\ of\ stocks\ covered^{t-7 \to t-2})_{i,t-1} + \beta_4 Sentiment_{i,j,t-1} \times ln(\#\ of\ stocks\ covered^{t-7 \to t-2})_{i,t-1} + \delta \cdot X_{j,t-1} + User_i + Year\text{-}month_t + \varepsilon_{i,j,t}, \quad (2)$$

Where *i*, *j*, and *t* denote a user, a firm, and a month; $RET_{j,t}$ is the monthly stock return for firm *j* in month *t*, in excess of the three-month Treasury bill rate, and $Sentiment_{i,j,t}$ is the Social-media sentiment for firm *j* given by user *i* based on all tweets posted in month *t-1*. $Sentiment_{i,j,t}$ is standardized for a more convenient interpretation. $ln(\#\ of\ stocks\ covered^{t-7 \to t-2})_{i,t-1}$ is the natural logarithm of the number of unique stocks user *i* has tweeted about during the 6 months from *t-7* to *t-2*, and would replace by $ln(\#\ of\ industries\ covered^{t-7 \to t-2})_{i,t-1}$ in some specifications. The coefficient estimate's sign, magnitude, and significance on the interaction term, β_4, are of main interest. A positive (negative) and significant estimation of β_4 would suggest that the predictive accuracy of users' tweets over future stock returns improves (worsens) with stock or industry coverage.

Figure 8 below shows the timeline for computing stock returns, users' stock and industry coverage, and their social-media sentiment. The # of stocks covered (# of industries covered) is calculated by measuring the unique number of stocks a user mentioned in tweets from month *t-7* to *t-2*. *The sentiment* is calculated in month *t-1* to predict *Stock return* in month *t*.

Figure 8
Timeline For Measuring Stocks (Industries) Coverage, Sentiment and Stock Returns

Fig. 8 shows the timeline for measuring stocks (industries) coverage, sentiment and stock returns. Specifically, stocks (industries) coverage is measured over the period from *t-7* to *t-2;* social-media sentiment is calculated from posted tweets over the period in month *t-1;* stock returns in month *t* are to be explained.



$X_{j,t-1}$ is a vector of lagged firm controls, shown in previous literature to be significant in explaining future excess stock returns, including log of market capitalization *ln(ME)*, log trading volume *ln(VOL)*, log closing monthly stock prices *ln(PRC)*, past momentum returns calculated over the periods from months *t-3* to *t-2 RET(-2,-3)*, *t-6 to t-4 RET(-4,-6)*, and *t-12 to t-7 RET(-7,-12)*, log book-to-market equity ratio *ln(BM)* calculated as book equity of fiscal year ending in calendar year *t-1* divided by market capitalization at the end of December in year *t-1*, and dividend yield *Dividend yield* computed by summing all monthly dividends over the past 12 months divided by share price at month *t-1*. $User_i$ and $Year\text{-}month_t$ are the user and year-month fixed effects, respectively.

To reduce the effect of extreme observations on the estimation, observations with monthly excess returns exceeding 100% are trimmed. The dividend yields is also winsorized at the top 1% and the book-to-market equity ratios at the bottom 1% percentiles. Standard errors are double-clustered at the user and month levels. The estimation is qualitative similar if we cluster standard errors only at the month level. This finding is consistent with Thompson (2011), who shows analytically that bias to standard errors increases with the time-series variation in the variables of interest and with the ratio of cross-sectional units (users and firms) to time-series units (months). Clustering at the month level would be most effective and necessary to correct such bias.

A potential issue is that persistent user characteristics such as innate ability, IQ, gender, investment orientation, and trading styles could shape the way users view certain stocks, thus affecting the sentiment value they assign to those stocks. For example, Grinblatt et al. (2012) and Grinblatt et al. (2015) noted that individuals with a higher IQ trade better and are smarter at picking stocks, timing order executions, and being better at avoiding funds with high management fees. Moreover, Clement et al. (2007) showed that innate ability is persistent and can significantly determine success in analyst forecasts. Interestingly, Barber and Odean (2001) revealed the significant role of gender difference in stock investment by showing that men trade more aggressively than women due to overconfidence. Another important time-invariant user characteristic is a trading style that is heterogeneous across individuals. For instance, a user could trade on firm fundamentals or technical indicators, or both.

To account for these potentially confounding factors, user fixed effects are controlled for in the model. Besides, the inclusion of user fixed effects also helps alleviate potential sample-

selection bias (Kyriazidou, 1997). The decision to follow stocks is likely a function of users' personal preference, psychological traits, trading experiences, or even cognitive capacity. As such, all between variations in user heterogeneity is eliminated.

The identification of the relationship in question relies only on within-user variation but not within-firm variation, hence firm fixed effect is not used. In other words, the tests examine whether a within-user increase in portfolio complexity (i.e., stock or industry coverage) over time determines the predictability of social-media sentiment over future stock returns but not within-firm characteristics determine the prediction of future returns. The user-level methodology is new and cannot be found in prior literature. Nevertheless, I have confirmed this with several senior econometricians and researchers from my department and other institutions. More details of our discussions are available on request.

Straightly speaking, time-invariant user characteristics also affect the modelling choice between fixed-effect and random-effect models. Consider equation (2) again, if a random-effect model is used, it requires the covariance of user or year-month unobserved heterogeneity with all other independent variables to be zero, that is:

$$Cov(User_i, Z_{i,j,t}) = 0$$

$$Cov(Year\text{-}month_i, Z_{i,j,t}) = 0$$

Where $Z_{i,j,t} = \{Sentiment_{i,j,t}, ln(\#\ of\ stocks\ covered)_{i,t}, Sentiment_{i,j,t-1} \times ln(\#\ of\ stocks\ covered)_{i,t-1}, X_{j,t}\}$.

However, as discussed above, the user time-invariant unobserved heterogeneity (for example, IQ) is correlated with *Sentiment*. In the same vein, any macro-economic shock within the year-month will also affect the user's investment decision and the firm's operating performance. Fixed-effect modelling eliminates the covariance correlation by removing both $User_i$ and $Year\text{-}month_i$ terms together, and hence this is the preferred method.

TABLE 5
Social-Media Sentiment, Stock Coverage, and Stock Returns

This table reports results from baseline regressions examining the relation between social-media sentiment, stock (industry) coverage, and stock returns. The dependent variable is monthly excess stock returns (RET). *Sentiment* is our monthly measure of social-media sentiment, estimated using machine learning techniques and tweets in month *t*-1. *ln(# of stocks covered)* [*ln(# of industries covered)*] is the natural log of the number of stocks (industries) covered by user *i* over a six-month period from *t*-7 to *t*-2. Lagged firm controls include: natural log of market capitalization (*ln(ME)*), natural log of trading volume (*ln(VOL)*), natural log of stock prices (*ln(PRC)*); past momentum returns over months *j*-2 to *j*-3 (*RET(-2, -3)*), *j*-4 to *j*-6 (*RET(-4, -6)*), and *j*-7 to *j*-12 (*RET(-7, -12)*), dividend yield (*Dividend yield*), and natural log of the book-to-market equity ratios (ln(BM)). User and month fixed effects are included in all models. Standard errors are double-clustered at the user and month levels. *T*-statistics are reported in parentheses. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Dependent variable: *RET*				
	(1)	(2)	(3)	(4)	(5)
Sentiment	0.267***	0.826***	0.544***	0.866***	0.573***
	(2.799)	(3.535)	(2.836)	(3.585)	(2.874)
ln(# of stocks covered)		0.097***	-0.061*		
		(2.603)	(-1.842)		
Sentiment × ln(# of stocks covered)		-0.125***	-0.072***		
		(-3.425)	(-2.626)		
ln(# of industries covered)				0.125**	-0.074*
				(2.522)	(-1.763)
Sentiment × ln(# of industries covered)				-0.170***	-0.100***
				(-3.525)	(-2.702)
ln(ME)	1.949***	2.107***	1.945***	2.107***	1.944***
	(6.426)	(6.535)	(6.435)	(6.539)	(6.434)
ln(VOL)	-0.163	-0.15	-0.164	-0.15	-0.164
	(-0.806)	(-0.696)	(-0.812)	(-0.696)	(-0.811)
ln(PRC)	-1.811***	-1.922***	-1.807***	-1.925***	-1.807***
	(-5.483)	(-5.656)	(-5.486)	(-5.675)	(-5.486)
RET(-2,-3)	-1.016***	-1.007***	-1.011***	-1.009***	-1.011***
	(-2.908)	(-2.850)	(-2.903)	(-2.853)	(-2.903)
RET(-4,-6)	0.018**	0.019**	0.018**	0.019**	0.018**
	(2.293)	(2.282)	(2.297)	(2.283)	(2.297)
RET(-7,-12)	0.018**	0.019**	0.018**	0.019**	0.018**
	(1.966)	(1.975)	(1.967)	(1.976)	(1.966)
Dividend yield	0.013***	0.016***	0.013***	0.016***	0.013***
	(2.814)	(3.132)	(2.814)	(3.134)	(2.813)
ln(BM)	-0.399	-0.459	-0.397	-0.459	-0.397
	(-1.458)	(-1.587)	(-1.451)	(-1.587)	(-1.451)
User FE	Yes		Yes		Yes
Year-month FE	Yes	Yes	Yes	Yes	Yes
Observations	5,556,568	5,556,568	5,556,568	5,556,568	5,556,568
Adjusted R^2	0.144	0.125	0.144	0.125	0.144

Table 5 reports the estimation results. Now the question is whether social-media sentiment can significantly predict future stock returns on average. In column (1), where *Sentiment*, firm controls, and user and month fixed effects are included, the social-media sentiment correlates positively and significantly (at the 1% level) to explain future stock returns. A one-standard-deviation increase in *Sentiment* is associated with an increase in excess returns of 26.7 basis points per month. My evidence is in line with recent studies on social media (see, e.g., Renault, 2017; Bartov et al., 2018).[23]

For firm controls, firms in the sample where higher market capitalization predicts higher next month returns, a one percent increase in *ln(ME)* predicts a 1.949% increase in a future return. Higher *Dividend Yield* also indicates a significant future return by 0.013. Interestingly, the higher price of stocks *ln(PRC)* does not predict higher future returns. A one percent increase predicts significantly -1.811% future return. The negative association on return with the price could suggest users primarily focus on small stocks in the sample used.

Furthermore, the momentum factor from short-term *RET(-2, -3)* shows a significant negative association with the future return, suggesting a trend reversal. However, factors like *RET(-4, -6)* and *RET(-7, -12)* significantly predict a positive future return in longer-term momentum. The short to longer-term momentum factor shows consistency with Hong and Stein (1999) model, where the early trader enters the market and holds a higher premium for an extended period. In contrast, investors joining late makes negative future return.[24] All firm controls remain consistent with signs across different specifications when user, year-month fixed effects are sequentially included in the regression.

[23] Bartov et al. (2018) documented evidence based on Twitter data that sentiment aggregated over the previous 9 days significantly and positively predicts buy-and-hold returns over the next three days. Likewise, Renault (2017) estimates sentiment in half-hour intervals using StockTwits data, finding that changes in social-media sentiment between the first half-hour interval in day *t* and the last half-hour interval in day *t*-1 predict positively and significantly market returns over the last half-hour interval in day *t*.

[24] The sign of the coefficient estimates for log market capitalization (*ln(ME)*) and log book-to-market equity ratio (*ln(BM)*) is inconsistent with prior asset-pricing studies for two main reasons. First, while prior studies use a stock-month dataset, our tests are estimated on a user-firm-month panel dataset. Firms that receive more interest from users and are thus being tweeted more would have more observations than others. As such, our estimates may be driven by the over-representation of these firms in our sample. Second, while prior asset-pricing studies examine the cross-sectional relationship between stock returns, size, and book-to-market equity ratios, our estimation includes user fixed effects and thus relies on the time-series, within-user variation in size, value, and stock returns for identification. Nonetheless, in unreported tests, we examine the relationship of size and book-to-market equity ratios with stock returns at the stock-month level, finding that the signs are consistent with prior studies.

Columns (2)-(5) report estimation results for equation (2), in column (2), where firm controls and year-month fixed effects are included. Including year-month fixed-effect further controls any macroeconomic shocks in the month that could affect the user's Sentiment and is baseline across specifications. The interaction term between social-media sentiment and stock coverage (*Sentiment × ln(# of stocks covered)*) enters negatively and significantly (at the 1% level) into the model. Further controlling for user fixed effects, column (3) shows that the negative coefficient estimate for the interaction term is significant at the 1% level, despite being smaller in magnitude. To gauge the economic magnitude, when *# of stocks covered* is at the 25^{th} [75^{th}] percentile, a one-standard-deviation increase in *Sentiment* is associated with future stock returns of (*ln(10.0)×-0.072 + 0.544=0.00378*) 37.8 basis points per month *[(ln(180.0)×-0.072+0.544=0.00170)* 17.0 basis points per month].

Columns (4) and (5) present results for regressions replacing *ln(# of stocks covered)* with *ln(# of industries covered).* Results show a similar pattern. In both columns, I find that the coefficient estimates for *Sentiment × ln(# of industries covered)* is again negative and highly significant (at the 1% level). Economically speaking, when the number of industries covered is at the 25^{th} and 75^{th} percentiles, a one-standard-deviation increase in *Sentiment* predicts higher future stock returns of (*ln(7.0)×-0.10+0.573=0.00378*) 37.8 basis points and [(*ln(67.0)×-0.10+0.753=0.00153)*] 15.3 basis points per month] 15.3 basis points per month, respectively.

Overall, these results reveal that users' sentiment following a larger set of stocks is worse at predicting future stock returns.

4.4.2 Robustness tests

This section provides several additional robustness tests for equation (2) and reports results in Table 6. To save space, the coefficients of interest is reported and the estimates is suppressed for firm controls.

Columns (1) and (2) apply alternative industry classifications for industry-coverage variables, including Fama-French 49-industry and 4-digit SIC industry classifications. Results are similar in magnitude and significance.

Columns (3) to (5) adjust stock returns for factor risk and industry effects. In column (3), adjust returns using the Carhart (1997) 4-factor model, estimated using returns over the past six years. Columns (4) and (5) remove median returns in Fama-French 49 industry and the 3-digit SIC industry. Results are robust to such risk and industry adjustments.

In columns (6) and (7), an alternative 3-month window is used for counting the unique number of stocks and industries covered by users, finding that the results continue to hold.

Thus far, the main tweet-sentiment measure is computed as the number of bullish tweets divided by the total number of tweets *within* a user-firm-month cell. As such, tweets posted on days on which the user has posted an unusually large number of tweets receive more weight in the estimation than vice versa. This approach is justified to the extent that users tend to tweet more after receiving new information, and new information does not arrive daily. For robustness, an alternative weighting scheme is applied by first calculating social-media sentiment daily then equal-averaging it across the days in a month; in effect, this gives equal weighting to each day in a month. As shown in columns (8) and (9), equally-weighted social-media sentiment results are qualitatively similar.

Columns (10) and (11) exclude illiquid or extremely large stocks where closing prices in month *t*-1 are below $1 or above $1,000. These stocks are less likely to be traded by individual investors because prices and volatilities tend to be high. After excluding these stocks, the results are still intact.

Lastly, since StockTwits users may have a stronger interest in covering technology stocks and thus post more tweets about them, the estimation results may be driven by the possible over-representation of such companies. I followed Barron et al. (2002) and remove technology stocks

based on their 3-digit SIC codes to address this concern.[25] As columns (12) and (13) show, while the number of observations drops consistent with suggestions where users focus heavily on technology stocks, results remain robust, meaning that this concern is unlikely to be significant.

[25] Technology firms are defined as those with a 3-digit SIC code as follows: 283, 284, 357, 366, 367, 371 382, 384, and 737.

TABLE 6
Robustness Tests

This table reports the results from our robustness tests. Columns (1) and (2) apply alternative industry classifications for the industry-coverage variables, including the Fama-French 49-industry and the 4-digit SIC industry classifications. Column (3) adjusts excess stock returns for factor risk using the Carhart (1997) 4-factor model. Columns (4) and (5) industry-adjust excess stock returns by de-medianing within the 49 Fama-French industries (FF49) and 3-digit SIC industries (SIC3). Columns (6) and (7) apply an alternative 3-month window for estimating the number of stocks and industries covered by a given user. Columns (8) and (9) apply an equal-weighting scheme for estimating social-media sentiment. Specifically, we first compute a daily measure of sentiment (total number of bullish tweets divided by the total number of tweets) for each user-firm pair and then average the daily measure within a month. Columns (10) and (11) use a sample that excludes stocks with prices smaller than $1 or larger than $1000. Following Barron et al. (2002), columns (12) and (13) exclude stocks operating in technology industries, defined as those having the following 3-digit SIC industry codes: 357, 737, 283, 382, 384, 366, 367, 284, or 371. Firm controls and fixed effects identical to the baseline models are included (their estimates are suppressed for brevity). Standard errors are double-clustered at the user and month levels. *T*-statistics are reported in parentheses. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Alternative industry classifications for industry coverage		Adjusting stock returns			Alternative window		Alternative measures of sentiment		Stock filter		Excluding Tech Stocks	
	FF49	SIC4	C-4	FF49	SIC3	3-month		EW		$1 < PRC < $1000		Exclude SIC3	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Sentiment	0.603***	0.569***	0.403**	0.592***	0.599***	0.516***	0.538***	0.544***	0.574***	0.420*	0.445*	0.752**	0.834***
	(2.856)	(2.841)	(1.968)	(3.145)	(3.178)	(2.807)	(2.838)	(2.857)	(2.896)	(1.943)	(1.951)	(2.567)	(2.606)
Sentiment × ln(# of stocks covered)			-0.061**	-0.079***	-0.080***	-0.071**		-0.072***		-0.058**		-0.064*	
			(-2.026)	(-2.892)	(-2.945)	(-2.566)		(-2.646)		(-2.010)		(-1.808)	
Sentiment × ln(# of industries covered)	-0.129***	-0.091***					-0.096***		-0.100***		-0.080**		-0.080*
	(-2.656)	(-2.649)					(-2.636)		(-2.724)		(-2.024)		(-1.725)
ln(# of stocks covered)			-0.047	-0.055*	-0.053*	-0.050*		-0.061*		-0.051**		-0.104**	
			(-1.568)	(-1.749)	(-1.730)	(-1.671)		(-1.842)		(-1.995)		(-2.507)	
ln(# of industries covered)	-0.118**	-0.077**					-0.064*		-0.074*		-0.056*		-0.154**
	(-2.246)	(-1.987)					(-1.708)		(-1.763)		(-1.807)		(-2.576)
Firm Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,556,568	5,556,568	4,042,140	5,556,638	5,557,163	5,520,877	5,520,877	5,556,568	5,556,568	5,069,489	5,069,489	2,832,528	2,832,528
Adjusted R^2	0.144	0.144	0.089	0.088	0.087	0.145	0.145	0.144	0.144	0.143	0.143	0.21	0.21

4.4.3 Subsample tests by stock and industry coverage

To enhance the creditability of the result, a subsample test is provided on how return predictivity of social media sentiment varies with the number of stocks and industries covered by user. Following regression specification on equation (2), which includes user and year-month fixed effect, subsample tests are performed that divide the sample into different user groups according to their stock coverage (0-20, 21-40, 41-60, 61-80, 81-100, 100+) and run regression on each bin.

TABLE 7
Baseline Tests By Users Groups Based On Stock Coverage

This table reports subsample results from baseline regressions (Table 3) examining the relation between social-media sentiment, stock coverage, and future stock returns in different stock-coverage categories. Our sample is divided into 6 groups according to the *# of stocks covered* at an increment of 20. The dependent variable is monthly excess stock returns (*RET*). *Sentiment* is our monthly measure of social-media sentiment, estimated using machine learning techniques and tweets in month *t*-1. *ln(# of stocks covered)* [*ln(# of industries covered)*] is the natural log of the number of stocks (industries) covered by user *i* over a six-month period from *t*-7 to *t*-2. Lagged firm controls include: natural log of market capitalization (*ln(ME)*), natural log of trading volume (*ln(VOL)*), natural log of stock prices (*ln(PRC)*); past momentum returns over months *j*-2 to *j*-3 (*RET(-2, -3)*), *j*-4 to *j*-6 (*RET(-4, -6)*), and *j*-7 to *j*-12 (*RET(-7, -12)*), dividend yield (*Dividend yield*), and natural log of the book-to-market equity ratios (*ln(BM)*). User and month fixed effects are included in all models. Standard errors are double-clustered at the user and month levels. *T*-statistics are reported in parentheses. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Dependent variable: *RET*					
	# of stocks coverage					
	1-20	21-40	41-60	61-80	81-100	100+
	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment	0.350**	0.315***	0.286***	0.247***	0.194**	0.119**
	(2.468)	(2.889)	(2.942)	(2.758)	(2.417)	(2.243)
ln(# of stocks covered)	-0.200***	-0.251	-0.281	0.230	1.152	0.055
	(-4.165)	(-1.545)	(-0.792)	(0.385)	(1.178)	(1.169)
ln(ME)	2.464***	2.234***	1.932***	1.715***	1.573***	1.064***
	(6.694)	(6.445)	(5.907)	(5.798)	(5.847)	(5.368)
ln(VOL)	-2.410***	-2.079***	-1.762***	-1.598***	-1.378***	-0.897***
	(-5.586)	(-5.292)	(-4.926)	(-4.743)	(-4.796)	(-4.687)
ln(PRC)	-1.296***	-1.239***	-1.041***	-0.936***	-0.818***	-0.429**
	(-2.825)	(-3.019)	(-2.793)	(-2.868)	(-2.847)	(-2.350)
RET(-2,-3)	0.020**	0.020**	0.018**	0.015**	0.013*	0.010*
	(2.245)	(2.480)	(2.571)	(2.265)	(1.927)	(1.704)
RET(-4,-6)	0.018	0.019**	0.018**	0.017**	0.014**	0.014***
	(1.570)	(1.982)	(2.205)	(2.514)	(2.339)	(2.826)
RET(-7,-12)	0.016***	0.015***	0.012***	0.011**	0.008**	0.006*
	(2.939)	(2.755)	(2.578)	(2.425)	(2.048)	(1.701)
Dividend yield	-0.502	-0.367	-0.322	-0.275	-0.145	-0.153
	(-1.287)	(-1.179)	(-1.283)	(-1.145)	(-0.753)	(-1.031)
ln(BM)	-0.211	-0.192	-0.206	-0.276	-0.135	-0.066
	(-0.743)	(-0.827)	(-1.032)	(-1.479)	(-0.811)	(-0.564)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,151,247	671,743	368,119	249,397	177,356	1,938,706
Adjusted R^2	0.153	0.139	0.132	0.121	0.124	0.121

Table 7 reports the subsample results according to the specification that follows equation (2). Each column corresponds to one of the six groups, and stock coverage increases from the leftmost to the rightmost columns. By running separate regressions through subsampling, it essentially interacts all control variables with stock coverage. Note that subsampling into different users group and running regression could produce different answers. Firstly, errors in each regression will be different as compared to full sample regression. Secondly, results may not hold for a certain group but will hold in the full sample.

Consistent with the baseline results, the coefficient for *Sentiment* is positive and significant at the 5% level or better in all columns. Notably, the estimates increase monotonically from the smallest (column 1) to the largest (column 6) stock-coverage groups, ranging from 0.35 to 0.12, entirely consistent with our baseline results. Regarding economic significance, for the group with 20 or fewer stocks covered, a one standard deviation increase in *Sentiment* is associated with a 35-basis-point increase in future excess stock returns. For an average user who follows more than 100 stocks, Sentiment's coefficient of one standard deviation increase is associated with an 11.9-basis-point increase in future excess stock returns.

The smallest (column 1) stock-coverage groups reveal a one percent increase in log market capitalization *ln(ME)* predicts the highest future return of 2.464% compared to the high coverage group (column 2) of 1.064%, all significant at 5% level or higher. Banz (1981) documented that firms with a low market capitalization have a higher return for the small firm effect. Results suggest low group users' firm covered predicts higher future return than others and, implying that low coverage group users follow more, on average, small firm's stocks. Volume associated with stock liquidity, Lee and Swaminathan (2000) found a positive relationship between volume and future returns. Here, log trading volume predicts -2.410% (column 1) in the lowest coverage group lower than the highest coverage group of -0.897% (column 6) and is significant at 1%. Based on the interpretation that an increase in trading volume predicts lower future return in both results, one shall see that the lower the user stock's coverage, the more negative return prediction for more volume traded.

4.4.4 Additional controls

This section further controls several additional users and tweet characteristics that could confound the results, Table 8 below reports the results.

Firstly, users gain reputation and fame through sharing profitable trading ideas and accurate price predictions. To show that results are independent of user reputation, in columns (1) and (2) highlight the interaction between *Sentiment* and the log of the number of likes received from users from month *t*-7 to *t*-2 (*ln(# of likes)*) in equation (2). Consistent with the expectation that more reputable users give more accurate price predictions, I find a positive and significant interaction term between social-media sentiment and the number of likes received. However, such an effect is independent of the baseline results. The estimates for the interaction terms between social-media sentiment and stock and industry coverage remain negative and highly significant.

Secondly, as shown in Table 1, users who follow fewer stocks post more tweets on average than those following more stocks. To show that stock coverage is not simply capturing users being more active in posting tweets and interacting with other users, we augment equation (2) with an additional interaction term between social-media sentiment and the natural log of the total number of tweets posted by a given user *i* about firm *j* in month *t-1*. As shown in columns (3) and (4), the interaction term between social-media sentiment and the number of tweets posted enters positively and significantly, consistent with the view that users who are more active in posting tweets and engaging with users better predict future stock returns. Our baseline results continue to hold.

Thirdly, the amount of effort and attention needed to research or analyse a stock is likely to be inversely related to the users' experience and previous time spent on it. If users persistently follow stocks with which they have had previous experience or are familiar, their significant predictive ability may stem from superior knowledge about the company's development. To control for users' past experience with a given stock, the natural log of the number of months are computed since a stock was first tweeted about by a user (*ln(# of months since the first tweet)*) and interact it with social-media sentiment to explain future stock returns. As shown in columns (5) and (6), our results remain qualitatively similar.

Finally, previous research suggests that companies in which analysts have more disagreements over earnings have lower future stock returns (see, e.g., Yu, 2011; Hong and Sraer, 2016). Since analyst recommendations are available to the general public, baseline results may be driven by the possibility that StockTwits users mainly follow analysts' views. As such, the

accuracy of social-media sentiment may be lower for firms with high analyst disagreement. Measuring analyst disagreement using I/B/E/S earnings forecast data, following Yu (2011), this is not a concern as I find the interaction term of interest (*Sentiment* × *ln(# of stocks covered)*) continues to hold after controlling for the interaction between social-media sentiment and disagreement.

TABLE 8
Additional Controls

This table presents the baseline tests with several additional user and firm characteristics as controls. The dependent variable is monthly excess stock returns (*RET*). Sentiment is our monthly measure of social-media sentiment, estimated using machine learning techniques and tweets in month *t*-1. *ln(# of stocks covered)* [*ln(# of industries covered)*] is the natural log of the number of stocks (industries) covered by user *i* over a six-month period from *t*-7 to *t*-2. *# of likes* is the number of likes received by a given user at the end of the six-month period from *t*-7 to *t*-2. *# of tweets* is the total number of tweets posted by a user for a firm in month *t*-1. *ln*(*# of months since the first tweet)* is the natural log of the number of months since a stock was first tweeted about by a user up to month *t-1*. *Analyst disagreement* is the standard deviation of analyst forecasts of the earnings-per-share (*EPS*) long-term growth rate (*LTG*) for a given firm in month *t*-1, following Yu (2011). Firm controls and fixed effects identical to the baseline models are included. Standard errors are double-clustered at user and month levels. *T*-statistics are reported in parenthesis. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Dependent variable: *RET*							
	# of likes		*# of tweets*		*Past experience*		*Analyst disagreement*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sentiment	0.436***	0.466***	0.406***	0.431***	0.551***	0.581***	0.557***	0.573***
	(2.838)	(2.890)	(2.888)	(2.942)	(2.797)	(2.837)	(2.963)	(2.959)
ln(# of stocks covered)	-0.019		-0.058*		-0.053		-0.056*	
	(-0.649)		(-1.737)		(-1.620)		(-1.704)	
Sentiment × ln(# of stocks covered)	-0.076***		-0.065***		-0.071***		-0.072***	
	(-2.593)		(-2.643)		(-2.827)		(-2.667)	
ln(# of industries covered)		-0.021		-0.071*		-0.065		-0.069*
		(-0.596)		(-1.676)		(-1.581)		(-1.654)
Sentiment × ln(# of industries covered)		-0.105***		-0.090***		-0.099***		-0.094***
		(-2.673)		(-2.726)		(-2.950)		(-2.667)
ln(1 + # of likes)	-0.075**	-0.076**						
	(-2.202)	(-2.309)						
Sentiment × ln(1 + # of likes)	0.033**	0.034**						
	(1.977)	(1.990)						
ln(# of tweets)			-0.285***	-0.285***				
			(-3.877)	(-3.883)				
Sentiment × ln(# of tweets)			0.314***	0.313**				
			(2.577)	(2.573)				
ln(# of months since the first tweet)					-0.063	-0.064		
					(-1.011)	(-1.027)		
Sentiment. × ln(# of months since the first tweet)					-0.008	-0.009		
					(-0.210)	(-0.228)		
Analyst disagreement							0.498***	0.497***
							(3.470)	(3.469)
Sentiment × Analyst disagreement							-0.338***	-0.336***
							(-3.808)	(-3.804)
Firm Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,556,568	5,556,568	5,556,568	5,556,568	5,556,568	5,556,568	5,556,568	5,556,568
Adjusted R^2	0.144	0.144	0.144	0.144	0.144	0.144	0.145	0.145

4.5 Investigating the Economic Mechanisms

Section 4.4.1 shows that tweets posted by users who followed fewer stocks or industries over the past 6 months have more substantial predictive power over future stock returns. This finding is consistent with the view that human beings have limited cognitive capacity and attention span. Thus, they can only allocate limited time, effort, and attention to analysing, researching, and studying stocks on their "watch list". As the number of followed stocks increases, the amount of time and attention paid to each decrease. Since the formulation of trading ideas and views requires some research, for example, analysing financial statements, company news, industry and market trends, etc. Users following a large number of stocks may simultaneously be less familiar with and less informed about the companies. Their stock-price predictions and recommendation are likely to be less accurate compared to users who follow only a small set of stocks and thus have more time and attention.

In the following sections, results are presented to examine the validity of this limited-attention explanation. First, the baseline test is re-estimated by excluding users whose stock price predictions are less likely to analyse and study fundamentals, company news, and markets. Second, focus on the ability of users to predict company earnings and whether such ability varies with stock and industry coverage. Third, explore the cross-sectional differences across firm characteristics that capture the degree of complexity and the ease with which users can research these companies.

4.5.1 Excluding users whose attention is less likely to be relevant

Limited attention explanation requires that an average user spend time and attention in studying and analysing the stocks they follow. While the analysis thus far has controlled for user fixed effects, this interpretation based on limited attention may not apply to all users, especially to those for whom analysing and researching company fundamentals and industry characteristics is less necessary for generating trading ideas or views. Examples of such users rely mainly on technical trading rules or those who trade with very short horizons, such as day traders. To show support for limited attention explanation, these users are excluded from our baseline regression, focusing on users who likely expend effort in researching companies and their fundamentals.

TABLE 9
Could Our Results Be Driven By Technical Traders or Other Noisy Users?

This table presents the tests addressing whether our results are driven by certain groups of users who are less likely to rely on firm fundamentals in formulating their investment views, decisions, or recommendations. In each row, we apply additional filters based on StockTwits users' self-reported characteristics, estimate the baseline tests on the filtered sample, and report only the coefficient estimates for *Sentiment*, *ln(# of stocks covered)*, and their interaction term. The rightmost three columns report the estimation based on the industry coverage variable. Row (1) excludes those users reporting that they have "professional" experience in investment and trading. Row (2) excludes users whose self-reported investment approach is "technical". Row (3) excludes those users self-reported as "day trader." Row (4) applies filters based on investment horizons and excludes all users except those stating that they are "long-term" investors. Row (5) applies the filters in rows (1), (2), and (3) simultaneously. Firm controls and fixed effects identical to the baseline models are included and suppressed for brevity. Standard errors are double-clustered at the user and month levels. *T*-statistics are reported in parenthesis. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

		Stocks coverage			Industries coverage		
Row	Excluding	*Sentiment*	*ln(# of stocks covered)*	*Sentiment* × *ln(# of stocks covered)*	*Sentiment*	*ln(# of industries covered)*	*Sentiment* × *ln(# of industries covered)*
(1)	Professional experience trader	0.557***	-0.083**	-0.069***	0.577***	-0.098**	-0.094***
		(2.886)	(-2.217)	(-2.599)	(2.909)	(-2.132)	(-2.664)
(2)	Technical trader	0.596***	-0.084**	-0.083***	0.612***	-0.107**	-0.109***
		(2.960)	(-2.152)	(-2.913)	(2.974)	(-2.206)	(-2.936)
(3)	Day trader	0.559***	-0.070**	-0.077***	0.585***	-0.084*	-0.105***
		(2.888)	(-1.990)	(-2.751)	(2.921)	(-1.906)	(-2.812)
(4)	Incl. only long-term investor	0.505***	-0.087**	-0.086***	0.542***	-0.106**	-0.118***
		(3.459)	(-2.506)	(-3.946)	(3.495)	(-2.355)	(-3.970)
(5)	(1) + (2) + (3)	0.605***	-0.103**	-0.082***	0.612***	-0.125**	-0.104***
		(3.001)	(-2.365)	(-2.940)	(3.005)	(-2.388)	(-2.936)

Table 9 presents results with additional filters. Row (1) excludes users who classify themselves as having professional experience as a trader. Row (2) excludes users who declare themselves "technical traders". Row (3) excludes users who identify themselves as a "day trader". Row (4) includes only users who consider themselves long-term investors. Row (5) includes the filters in rows (1) to (3) simultaneously. The main results in all rows continue to hold for both the stock- and industry-coverage variables, consistent with the limited-attention explanation.

4.5.2 Social-media sentiment, stock coverage, and earnings

To the extent that users following a small set of stocks predict stock returns more accurately, due to more time and attention allocated to studying the companies. It's expected to find that such users are also superior at predicting a company's fundamental performance, most notably, earnings. To test this, the following regression model is estimated:

$$EPS_{j,t} \text{ or } SUE_{j,t} = \beta_0 + \beta_1 Sentiment^{k-22 \rightarrow k-1}_{i,j,t} + \beta_2 ln(\#\ of\ stocks\ covered^{t-7 \rightarrow t-2})_{i,t-1} + \beta_1 Sentiment^{k-22 \rightarrow k-1}_{i,j,t} \times ln(\#\ of\ stocks\ covered^{t-7 \rightarrow t-2})_{i,t-1} + X_{j,t-1} + \varepsilon_{i,j,t}, \quad (3)$$

where i, j, k, and t denote a user, stock, an announcement day, and an announcement month. $EPS_{j,t}$ is the earnings per share for stock j in earning announcement month t. $SUE_{j,t}$ is the standardized unexpected earnings (SUE) of stock j in earning announcement month t, computed as current quarter earnings minus same quarter from last year's earnings, divided by stock price in the last quarter, following Ayer et al. (2011). [26]

Social-media sentiment ($Sentiment^{k-22 \rightarrow k-1}_{i,j,t}$) in this specification is defined as the total number of bullish tweets divided by the total number of tweets over the past 21 trading days, prior to an earnings announcement (skipping the day before the announcement to reduce potential microstructure effect). The stock and industry coverage variables are similarly computed by counting the number of stocks or industries tweeted about by a user over the period from t-7 to t-2. Similarly, $X_{j,t-1}$ is a vector of firm controls, including natural of log market capitalization ($ln(ME)$) on day k-2, natural log of daily trading volume ($ln(VOL)$) on day k-2, one-year lagged book-to-market equity ratios ($ln(BM)$), and daily abnormal returns estimated from the Carhart (1997) 4-factor model, using daily returns over the past 6 years up to day k-2.

[26] In unreported analysis, since the composition of SUE is the difference between this quarter and same quarter from last year's EPS and is considered to be a longer-term surprise measure, we use shorter-term earnings surprise measure calculated by the difference in EPS, and contemporaneous analyst median forecasts as dependent variables, finding even stronger statistical significance in the interaction term between tweet sentiment and the stock- and industry-coverage variables.

Moreover, one-quarter lagged earnings per share or *SUE*[27], following prior accounting literature (see, e.g., Bartov et al., 2018), is included in the number of analysts following the company in the month *t*-2, two dummy variables for the fiscal year-end and the fourth quarter, and the number of tweets posted by a user. User and month fixed effects are included in all models. Standard errors are double-clustered at the user and month levels.

TABLE 10
Social-Media Sentiment, Stock Coverage, and Earnings Predictability

This table reports the results from regressions examining the relation between social-media sentiment, stock coverage, and earnings predictability. The dependent variables are earnings per share (*EPS*), defined as current-quarter earnings, and standardized unexpected earnings (*SUE*), which is computed as current-quarter earnings minus same-quarter-from-last-year earnings, scaled by last-quarter stock price. *Sentiment* is the monthly social-media sentiment, estimated using machine learning techniques. *ln(# of stocks covered)* [*ln(# of industries covered)*] is the natural log of the number of stocks (industries) covered by user i over the six-month period from *t*-7 to *t*-2. In addition to the baseline firm controls, we include abnormal returns (*Abnormal RET*) estimated from the Carhart (1997) four-factor model, one-quarter-lagged *EPS* or *SUE*, natural log of the number of analysts following the firm, and dummy variables indicating fiscal year end, and a loss-making dummy from previous quarter. User and month fixed effects are included in all models. Columns (1) and (2) report results using *EPS* as dependent variable. Columns (3) and (4) show results with *SUE* as dependent variable. Standard errors are double-clustered at the user and month levels. *T*-statistics are reported in parenthesis. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	EPS		*SUE*	
	(1)	(2)	(3)	(4)
Sentiment	0.013***	0.014***	0.027**	0.030**
	(2.649)	(2.639)	(2.052)	(2.153)
ln(# of stocks covered)	0.002		0.006***	
	(1.389)		(2.694)	
Sentiment × ln(# of stocks covered)	-0.002***		-0.004*	
	(-3.174)		(-1.924)	
ln(# of industries covered)		0.002		0.007***
		(1.410)		(2.650)
Sentiment × ln(# of industries covered)		-0.003***		-0.006**
		(-3.049)		(-2.075)
ln(# of tweets)	0.002	0.002	-0.001	-0.001
	(0.581)	(0.581)	(-0.274)	(-0.271)
Sentiment × ln(# of tweets)	0.011***	0.011***	0.021**	0.021**
	(4.483)	(4.464)	(1.994)	(1.990)
ln(ME)	0.080***	0.080***	0.019*	0.019*
	(7.475)	(7.474)	(1.955)	(1.954)
ln(VOL)	-0.048***	-0.048***	0.004	0.004
	(-5.256)	(-5.256)	(0.219)	(0.219)
ln(BM)	0.015*	0.015*	-0.037**	-0.037**
	(1.854)	(1.856)	(-2.279)	(-2.280)
Abnormal RET	0.001	0.001	0.007**	0.007**
	(1.261)	(1.260)	(2.535)	(2.536)
Lagged EPS	0.779***	0.779***		
	(36.284)	(36.283)		
Lagged SUE			0.331***	0.331***
			(7.746)	(7.746)
# of analyst	-0.001	-0.001	-0.004***	-0.004***
	(-0.788)	(-0.788)	(-3.290)	(-3.291)
Fiscal year end dummy	-0.063	-0.063	0.007	0.007
	(-1.633)	(-1.633)	(0.437)	(0.436)
Last quarter loss dummy	0.087***	0.087***	0.054	0.054
	(3.014)	(3.014)	(1.354)	(1.353)
User FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	542,411	542,411	517,823	517,823
Adjusted R^2	0.785	0.785	0.283	0.283

[27] Noted that lagged dependent variable such as *Lagged SUE* will lead to Nickell (1981) bias in fixed effect model and this is one of the limitation in this analysis. This means when demeaning fixed effect model the demeaned *Lagged SUE* leads to correlation with the demeaned error term because error term contains *Lagged SUE*. Nickell (1981) shows the bias is important when number of observation *N* is large and time *T* is small. However, consistent result is found by removing the control term *Lagged SUE*.

Estimation results for equation (3) are reported in Table 10. Columns (1)-(2) and (3)-(4) report results for *EPS* and *SUE*, respectively. These results are in line with our conjecture. In column (1), the coefficient for the interaction term between social-media sentiment and stock coverage is negative and highly significant. Economically speaking, when # of stocks covered is at the 25th percentile [75th percentile], a one-standard-deviation increase in *Sentiment* is associated with a $0.008 [$0.001] increase in EPS. It's expected *Lagged EPS* to have significant predictive power in the current quarter EPS for the control variables. Consistent with my prediction, lagged EPS can predict $0.779 of current quarter EPS and is a significant 1% level or more. *Last quarter loss dummy* is also positively significant, as with Bartov et al. (2018), which explains that the current quarter EPS will rise by $0.087 and again significant at 1% level if the previous quarter has a loss.

As in column (2), results using industry coverage are similar in both magnitude and significance. In column (3), where *SUE* is the dependent variable, when # of stocks covered is at 25th [75th percentile], a one-standard-deviation increase in *Sentiment* is 0.016 [0.004] of *SUE*. My result shows the *# of analyst* lowers *SUE* by -0.004 and is significant at 1% level. Consistent with prior accounting literature, the higher the number of analysts covering a stock, the fewer the unexpected earnings. Firms with a high log book-to-market ratio *ln(BM)* tend to lower the SUE by -0.037 and significantly at a 5% level. The fall in *SUE* by *ln(BM)* suggests value firms can minimize unexpected earnings compared to more volatile growth firms.

Furthermore, consistent with Bernard and Thomas (1990) in column (2), *Lagged SUE* is a significant predictor to current quarter *SUE* in quarterly earnings surprises. A one-unit increase of *Lagged SUE* predicts positively 0.331 of current quarter *SUE* and significant at 1% level. *Abnormal RET* captures information other than through StockTwits that may have reached the stock market before any earning announcement. My result consistently shows a significant positive relationship at the 5% level because it reflects earnings information that is not included in the current *SUE*.

The robustness of my results with filters section is checked to exclude all users who identify themselves as either professional traders, technical traders, or day traders. The result continues to hold, and the negative interaction terms of interest remain significant at the 10% level or better.

Overall, they are consistent with the limited-attention explanation, my findings suggest that tweets from users following fewer stocks can predict earnings and earnings surprises more accurately.

4.5.3 Exploring heterogeneity in stock characteristics

Suppose that the predictability of social-media sentiment over future stock returns declines with stock coverage due to reduced time and attention spent studying the companies. In that case, it is expected that the negative moderating effect of stock coverage will be more pronounced for more complex firms, hard to value, and where information is more asymmetric, opaque, and difficult to understand. The loss in accuracy in price predictions is likely to be most severe among such stocks when users' research effort is reduced due to behavioural constraints. To test this conjecture, the firms are divided into high and low groups according to several firm characteristics relating to firm complexity, information asymmetry, and uncertainty and estimate baseline tests from these subsamples.

Six empirical proxies, including firm *discretionary accruals*, *firm complexity*, *idiosyncratic volatility*, *firm age*, *analyst disagreement*, and *earnings volatilities*.[28] *Discretionary accruals* are considered to capture the abnormal proportion of companies' accruals to obscure information relating to fundamental position and thus measure information opaqueness (Sloan, 1996; Hutton et al., 2009). To measure *Discretionary accruals*, the difference between the actual and the "normal" accruals are computed in a given firm year. "Normal" accrual is the model-fitted values from estimating the modified Jones model (Dechow et al., 1995) for each 3-digit SIC industry and in each year. *Idiosyncratic volatility* is the standard deviation of residuals estimated using the Carhart (1997) 4-factor model and monthly stock returns over the past six years. Zhang (2010) showed that idiosyncratic volatility is driven by uncertainty in firm fundamentals, especially current and future earnings growth.

According to Cohen and Lou (2011), for firm complexity, companies with conglomerate multiple operating segments are more susceptible to macroeconomic shocks and harder to analyse due to the differing degree of information diffusion in each segment. Hence, they experience more sluggish price updating compared to standalone firms. Following Cohen and Lou (2011), standalone firms (Complicated firms) are defined as those operating in only one

[28] In unreported result, we also used cash flow volatility as proxy for firm opacity characteristics (Zhang (2006)) and results are in line with our prediction.

(greater than one) industry segment, and its segment (all segment) sales account for at least 80% of total sales. As for *Firm age*, younger firms tend to have less track record, higher uncertainty, and less informational transparency, and, hence, they are harder to value in general (Baker and Wrugler, 2006; Zhang, 2006). *Analyst disagreement*, measured as the standard deviation in analyst earnings per share (EPS) long-term growth rate (LTG) from the I/B/E/S database (Diether et al., 2002; Yu, 2011), captures the level of uncertainty surrounding firms about future earnings and is a proxy for information asymmetry. The final proxy is the standard deviation in return on assets (*Operating profitability volatility*) for the past 5 years (requiring at least 3 years of data for the estimation). Firms with more uncertain cash flows are harder to analyse and value. For each subsample test, the firms are divided by month according to the sample median of the above firm characteristics.[29]

[29] In the subsample tests based on *Analyst disagreement*, the firms are divided into high and low groups at the 80th percentile due to high skew. Results continue to hold if the firms are divided at the 75th percentile instead. As for *Firm complxity*, the low (high) group contains standalone firms that only operate in one segment (multiple segments).

TABLE 11
Subsample Tests

This table presents the results from subsample tests of five firm-level measures of the relative ease with which firms can be analysed and researched. Panels A and B report results based on the stock coverage and industry coverage variables, respectively. The dependent variable is monthly excess stock returns (*RET*). *Sentiment* is our monthly measure of social-media sentiment, estimated using machine learning techniques and tweets in month *t-1*. *ln(# of stocks covered)* [*ln(# of industries covered)*] is the natural log of the number of stocks (industries) covered by user *i* over a six-month period from *t-7* to *t-2*. The subsample tests are estimated at the user-firm-month level. *Disc. accruals* is the discretionary accruals from last fiscal year, computed as the difference between actual and "normal" accruals in a given firm and given year. Normal accrual is the model-fitted accruals, using estimated parameters from the modified Jones model (Dechow et al., 1995) in each 3-digit SIC industry and year. *Firm complexity* is a dummy that equals one for firms who operate in more than one 2-digit SIC industry and whose aggregate sales from all reported industry segments account for more than 80% of its total sales. It equals zero for standalone firms who operate only in one industry and whose segment sales account for more than 80% of its total sales. *Idiosyncratic risk* is the estimated residuals from the estimation of Carhart (1997) four-factor model, analysing monthly returns over the past 6 years. *Firm age* is the number of months since a stock's first appearance in the CRSP database. *Analyst disagreement* is the standard deviation of analyst forecast of earnings-per-share (*EPS*) long-term growth rate (*LTG*) for a given firm in month *t-1*, following Yu (2011). *Operating profitability volatility* is the standard deviation of return on asset (ROA), estimated over the past 5 years (requiring at least 3 years of observations). *ln(# of tweets)* and its interaction with social-media sentiment is controlled for in each model. The variables used for subsampling are also controlled for in the models. Firm controls and fixed effects identical to the baseline model are included. For brevity, we only report the estimates for the main variables of interest and suppress those of the other controls. Standard errors are double-clustered at user and month levels. *T*-statistics are reported in parenthesis. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A. Stock coverage

	Dependent variable: *RET*											
	Disc. accruals		*Firm complexity*		*Idiosyncratic risk*		*Firm age*		*Analyst disagreement*		*Operating profitability volatility*	
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Sentiment	-0.012	0.614**	0.252*	0.601**	-0.005	0.719***	0.664***	0.051	0.301	0.448***	-0.025	0.507**
	(-0.080)	(2.180)	(1.764)	(2.476)	(-0.111)	(3.307)	(2.968)	(0.423)	(1.535)	(2.694)	(-0.242)	(2.528)
ln(# of stocks covered)	-0.047*	-0.05	-0.046	-0.112*	0.001	-0.092**	-0.07	-0.046*	-0.065	-0.069*	0.001	-0.053
	(-1.816)	(-0.993)	(-1.159)	(-1.798)	(0.084)	(-2.177)	(-1.316)	(-1.791)	(-1.437)	(-1.810)	(0.047)	(-1.262)
Sentiment × ln(# of stocks covered)	0.009	-0.085*	-0.032	-0.114***	0.003	-0.107***	-0.106***	-0.005	-0.042	-0.084***	0.011	-0.076**
	(0.356)	(-1.812)	(-1.325)	(-3.028)	(0.367)	(-2.983)	(-2.746)	(-0.224)	(-1.043)	(-2.969)	(0.566)	(-2.290)
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for *Sentiment × ln(# of tweets)*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for dividing variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,392,153	2,404,975	3,843,060	808,315	1,365,998	2,632,660	2,621,100	2,901,302	2,198,160	3,324,242	1,575,309	3,642,414
Adjusted R^2	0.19	0.169	0.132	0.428	0.229	0.182	0.168	0.14	0.151	0.162	0.166	0.145
H_0: Coefficient equality for *Sentiment × ln(# of stocks covered)* across subsamples [*t*-stat]	-1.606		-1.951*		-2.983***		2.318**		-0.705		-2.268**	

Panel B. Industry coverage

	Dependent variable: *RET*											
	Disc. accruals		*Firm complexity*		*Idiosyncratic risk*		*Firm age*		*Analyst disagreement*		*Operating profitability volatility*	
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Sentiment	-0.007	0.620**	0.259*	0.687***	-0.015	0.745***	0.689***	0.057	0.322	0.473***	-0.034	0.512**
	(-0.045)	(2.116)	(1.743)	(2.612)	(-0.300)	(3.340)	(3.026)	(0.440)	(1.526)	(2.719)	(-0.281)	(2.525)
ln(# of industries covered)	-0.064*	-0.068	-0.053	-0.153*	-0.001	-0.112**	-0.085	-0.059*	-0.09	-0.079*	0.002	-0.072
	(-1.936)	(-1.074)	(-1.074)	(-1.786)	(-0.038)	(-2.096)	(-1.291)	(-1.754)	(-1.520)	(-1.769)	(0.092)	(-1.353)
Sentiment × ln(# of industries covered)	0.010	-0.108*	-0.043	-0.166***	0.007	-0.142***	-0.141***	-0.008	-0.059	-0.113***	0.016	-0.097**
	(0.270)	(-1.731)	(-1.320)	(-3.120)	(0.595)	(-3.024)	(-2.812)	(-0.263)	(-1.067)	(-2.988)	(0.565)	(-2.281)
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for *Sentiment × ln(# of tweets)*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for dividing variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,392,153	2,404,975	3,843,060	808,315	1,365,998	2,632,660	2,621,100	2,901,302	2,198,160	3,324,242	1,575,309	3,642,414
Adjusted R^2	0.19	0.169	0.132	0.428	0.229	0.182	0.168	0.14	0.151	0.162	0.166	0.145
H_0: Coefficient equality for *Sentiment × ln(# of industries covered)* across subsamples [*t*-stat]	-1.459		-2.063**		-3.059***		2.317**		-0.651		-2.207**	

Table 11 reports these subsample results. Panels A and B report results for stock coverage and industry coverage, respectively. All models include the baseline controls, fixed effects, and interaction between social-media sentiment and the number of tweets posted (*ln(# of tweets)*). Although suppressed for brevity, all models include the dividing variables to control for their direct effect on future stock returns.

In both panels, results reveal a clear pattern across the six firm complexity and information opaqueness proxies. *Discretionary Accrual* reflects the degree of the obscurity of a company's fundamental position. The interaction term in *Sentiment* × *ln(# of companies covered)* is negative and only significant in the "High" side. This indicates that users have difficulty analysing companies with high fundamental obscurity and large stock coverage to analyse, leading to lower future return predictions. The same goes for *Firm complexity* when a firm consists of several segments, as shown by the significant interaction term in the "High" side. Users require more research about the complex relationships of different segments in a firm to conclude the fundamental position of a firm, leading to lower return prediction, especially for users with more stocks covered in their portfolio. For *firm age*, if a firm is young, for example, a start-up, they do not have much prior fundamental information, and hence their fundamental is opaque to the user. We shall see the interaction term is significant in the "Low" side of the dividing variable. Consistently in panels A and B, and even the coefficient equality is pointing evidence towards users find it hard to analyse the firm, especially with their higher stocks/industries coverage which predicts lower future return.

For firms with high *Idiosyncratic risk*, *Operating profitability volatility*, and *Analyst Disagreement*, the same interpretation can be made with interaction term being significantly negative only at the "High" side. This means users find it hard to understand the information opaqueness variables. *Idiosyncratic risk* measures the firm-specific risk faced by the company; *Operating profitability volatility* presents how volatile the cash flow stream of a firm is. And lastly, *Analyst Disagreement* captures the disagreement among analysts for the firm. This finding lends direct support to the limited-attention explanation.

In the result with filters section, results continue to hold even after performing the subsample tests, excluding all users who label themselves as professional, technical, or day traders. Researching company fundamentals is less important for generating trading ideas.

4.6 Trading Strategy

Evidence in previous sections all suggests social-media sentiment is positively and significantly associated with higher future stock returns and earnings. This predictability decreases with the number of stocks or industries followed by users and is consistent with behavioural models such as Hirshleifer et al. (2011). When investors focus on many subsets of publicly available earnings, less time and effort is spent on researching each of these stocks, which leads to stock-price predictions that deviate further from fundamental value. In this section, a trading strategy based on previous findings is designed to evaluate its profitability. The design of such a trading strategy is similar to the one proposed by Giannini et al. (2017).

The construction of the trading strategy is as follows. First, users who tweeted are divided into two groups according to their stock coverage for each firm. The low-coverage (high-coverage) group includes users covering 1 to 100 stocks (more than 100 stocks).[30] The social-media sentiment is then averaged within the two groups to compute the difference in average sentiment (referred to as "net sentiment") between the low- and high-coverage groups (Low-Minus-High). The stock has two monthly sentiment scores for the high- and low-coverage groups and a net sentiment score. Stocks with positive net sentiment and tweets posted by users with low coverage are more bullish than tweets posted by users with high coverage. Since low-coverage users are shown to predict returns more accurately, stocks with zero or positive net sentiment should outperform stocks with a negative net sentiment. Hence, at the beginning of month *t*, the trading strategy longs a portfolio of stocks with a zero or positive net sentiment and shorts a portfolio of stocks with a negative net sentiment, rebalancing monthly.

[30] Since the number of users in the 1-100 group is larger than that in the 100+ group (see Panel B of Table 1) in untabulated analysis, I test an alternative trading strategy that replaces the 1-100 group with a group consisting of users with stock coverage between 1 and 20 stocks (1-20 group). Using the same procedures in computing the average sentiment scores and the net scores for the sample stocks, a strategy that longs stocks with a positive or zero net sentiment score and shorts stocks with a negative net sentiment score continues to generate significant (at the 10% level) trading profits, despite the profits become smaller in magnitude.

TABLE 12
Trading Strategy

This table evaluates the profitability of a trading strategy that is designed to exploit the variation in social-media sentiment and stock coverage. For each firm at the beginning of each month, I divide users into low and high groups according to their stock coverage (1-100 vs. 100+), equal-average the sentiment for the two groups, and compute the net sentiment value (low minus high). A trading strategy that longs stocks with a positive or zero net sentiment and short stocks with a negative net sentiment is designed, rebalancing every month. Panel A reports the equally-weighted average returns for the long, short, and long-short portfolios. Columns (4) to (6) report results based on subsamples excluding users that have declared themselves "professional," "technical," and a "day trader." Columns (7) to (9) report results based on subsamples excluding stocks that operate in technology industries, defined as those having the following 3-digit SIC codes, after Barron et al. (2002): 357, 737, 283, 382, 384, 366, 367, 284, or 371. Panel B reports the trading alpha (in %) of the long, short, and long-short zero-cost spread portfolios, estimated by regressing their monthly returns on market, size, value, and momentum factors, based on the Carhart (1997) 4-factor model. Newey-West robust standard errors with 12 lags are reported in parentheses. *T*-statistics are reported in parenthesis. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A. Excess returns

	Full sample			Excluding professional, technical, and day trader			Excluding technology firms		
Stock covered group	[1-100] - [100+]			[1-100] - [100+]			[1-100] - [100+]		
Net sentiment	<0	≥0		<0	≥0		<0	≥0	
Portfolios	Short	Long	Long-Short	Short	Long	Long-Short	Short	Long	Long-Short
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Avg. excess returns	0.674	0.891**	0.217***	0.647	0.803*	0.156*	0.531	0.779*	0.248***
	(1.494)	(2.225)	(2.672)	(1.403)	(1.899)	(1.862)	(1.230)	(1.931)	(3.416)
Observations	89	89	89	89	89	89	89	89	89

Panel B. Carhart (1997) 4-factor model

	Full sample	Excluding professional, technical, and day trader	Excluding technology firms
	Alpha (%)	Alpha (%)	Alpha (%)
<0 (Short)	-0.576***	-0.643***	-0.623***
	(-3.623)	(-3.250)	(-4.698)
≥0 (Long)	-0.312**	-0.392**	-0.376**
	(-2.111)	(-2.459)	(-2.483)
Long-Short	0.264***	0.251***	0.247***
	(2.940)	(3.006)	(3.295)

Panel A of Table 12 reports average portfolio excess returns for long and short portfolios for the full sample (see columns 1 to 3). Results are also reported to be based on alternative samples in which social-media sentiment is estimated from restricted samples of tweets, excluding those posted by users classifying themselves as professional, technical, or day traders (see columns 4 to 6), and those relating to technology firms (see columns 7 to 9). *T*-statistics reported in parentheses are based on Newey-West corrected robust standard errors. In the full-sample tests, it is found that the long portfolio (with positive net sentiment) yields an average monthly excess return of 89.1 basis points, statistically significant at the 5% level (*t*-statistics=2.225), whereas the short portfolio (with negative net sentiment) has an average excess return of 67.4 basis points (*t*-statistics=1.494), consistent with our expectations. More importantly, the long-short zero-cost spread portfolio yields an average return of 21.7 basis points per month, statistically significant at the 5% level, suggesting that the trading strategy makes an annual profit of approximately 2.6 percentage points. The excess returns on the spread portfolios are smaller but remain marginally significant after excluding professional, technical, or day traders. For instance, the long-short portfolio excluding professional, technical, or day traders generates 15.6 basis points and is only significant at 10% level compared to full sample sorting. Motivated by the fact that StockTwits users who talk more about technology stocks, my Long-short portfolio is significant by removing these firms. It achieves similar returns of 24.8 basis points and is significant at the 1% level.

Panel B reports the trading alpha (in %) of the long, short, and long-short zero-spread portfolios, adjusted using the Carhart (1997) 4-factor model.[31] Results show that both the long and short portfolios yield a significantly negative trading alpha, significant at the 5% level or better, implying that both portfolios underperform the market. Nonetheless, the positive returns to the zero-cost spread portfolio persist and become even larger and more significant after accounting for market risk, size, value, and momentum effects (26.4 basis point monthly; *t*-statistics=2.940), confirming that such trading strategy is profitable.

[31] The pricing factors are downloaded from Professor Kenneth French's data library. We thank Professor French for making such data publicly available.

4.7 Results with filters and variable definitions

To increase confidence with the results, filters are applied in this section. These filters performs by excluding all users who label themselves as professional, technical, or day traders. These users are excluded because they are less likely to use company fundamentals for research and generate trading ideas. Table 13 shows the main result when the filter is applied in section 4.4 Table 5. Table 14 refers to the subsample test in section 4.5.3 Table 11. All coefficients of interest remain significant at 10% or more. Also, Table 15 provides a summary of variable used in this chapter.

TABLE 13
Social-Media Sentiment, Stock Coverage, and Earnings Predictability With Filters

This table reports results from regressions examining the relation between social-media sentiment, stock coverage, and earnings predictability, after excluding users who label themselves as "professional," "technical," or "day trader." The dependent variables are earnings per share (*EPS*), defined as current-quarter earnings and standardized unexpected earnings (*SUE*), is computed as current-quarter earnings minus same-quarter-from-last-year earnings, scaled by last-quarter stock price. *Sentiment* is the monthly social-media sentiment, estimated using machine learning techniques. *ln(# of stocks covered)* [*ln(# of industries covered)*] is the natural log of the number of stocks (industries) covered by user i over the six-month period from *t*-7 to *t*-2. In addition to the baseline firm controls, we include abnormal returns (*Abnormal RET*) estimated from the Carhart (1997) four-factor model, one-quarter-lagged *EPS* or *SUE*, natural log of the number of analysts following the firm, and dummy variables indicating fiscal year end, and a loss-making dummy from previous quarter. User and month fixed effects are included in all models. Columns (1) and (2) report results using *EPS* as dependent variable; columns (3) and (4) show results with *SUE* as dependent variable. Standard errors are double-clustered at the user and month levels. *T*-statistics are reported in parenthesis. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	EPS		*SUE*	
	(1)	(2)	(3)	(4)
Sentiment	0.011**	0.011**	0.030**	0.034**
	(2.258)	(2.141)	(2.109)	(2.226)
ln(# of stocks covered)	0.001		0.006	
	(0.309)		(1.622)	
Sentiment × ln(# of stocks covered)	-0.002**		-0.004*	
	(-2.152)		(-1.829)	
ln(# of industries covered)		0.001		0.006
		(0.528)		(1.552)
Sentiment × ln(# of industries covered)		-0.002*		-0.006**
		(-1.946)		(-2.012)
ln(# of tweets)	0.001	0.001	-0.003	-0.003
	(0.445)	(0.446)	(-0.736)	(-0.732)
Sentiment × ln(# of tweets)	0.012***	0.012***	0.029**	0.029**
	(3.472)	(3.478)	(2.028)	(2.023)
ln(ME)	0.074***	0.074***	0.011	0.011
	(5.857)	(5.857)	(0.920)	(0.920)
ln(VOL)	-0.039***	-0.039***	0.004	0.004
	(-3.571)	(-3.571)	(0.163)	(0.163)
ln(BM)	0.015	0.015	-0.049**	-0.049**
	(1.448)	(1.449)	(-2.336)	(-2.336)
Abnormal RET	0.001	0.001	0.008**	0.008**
	(1.186)	(1.186)	(2.309)	(2.310)
Lagged EPS	0.778***	0.778***		
	(29.303)	(29.303)		
Lagged SUE			0.313***	0.313***
			(6.603)	(6.603)
# of analyst	-0.001	-0.001	-0.004*	-0.004*
	(-0.515)	(-0.515)	(-1.955)	(-1.955)
Fiscal year end dummy	-0.05	-0.05	0.005	0.005
	(-1.126)	(-1.126)	(0.239)	(0.238)
Last quarter loss dummy	0.076**	0.076**	0.043	0.043
	(2.291)	(2.291)	(0.924)	(0.923)
User FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	279,448	279,448	266,042	266,042
Adjusted R^2	0.795	0.795	0.299	0.299

TABLE 14
Subsample Tests With Filters

This table presents subsample the test results for six firm-level measures of the relative ease with which firms can be analysed and researched, after excluding users who label themselves as "professional," "technical," or "day trader." Panels A and B report results based on stock coverage and industry coverage variables, respectively. The dependent variable is monthly excess stock returns (*RET*). *Sentiment* is our monthly measure of social-media sentiment, estimated using machine learning techniques and tweets in month *t-1*. *ln(# of stocks covered)* [*ln(# of industries covered)*] is the natural log of the number of stocks (industries) covered by user *i* over a six-month period from *t-7* to *t-2*. The subsample tests are estimated at the user-firm-month level. *Disc. accruals* is the discretionary accruals from last fiscal year, computed as the difference between actual and "normal" accruals in a given firm and given year. Normal accrual is the model-fitted accruals, using the estimated parameters from the modified Jones model (Dechow et al., 1995) in each 3-digit SIC industry and year. *Firm complexity* is a dummy that equals one for firms who operate in more than one 2-digit SIC industry, and whose aggregate sales from all reported industry segments account for more than 80% of its total sales. It equals zero for standalone firms who operate only in one industry and whose segment sales account for more than 80% of its total sales. *Idiosyncratic risk* is the estimated residuals from the estimation of Carhart (1997) four-factor model, analysing monthly returns over the past 6 years. *Firm age* is the number of months since a stock's first appearance in the CRSP database. *Analyst disagreement* is the standard deviation of analyst forecast of earnings-per-share (*EPS*) long-term growth rate (*LTG*) for a given firm in month *t-1*, following Yu (2011). *Operating profitability volatility* is the standard deviation of return on asset (ROA), estimated over the past 5 years (requiring at least 3 observations). *ln(# of tweets)* and its interaction with tweet sentiment are controlled for in each model. The variables used for subsampling are also controlled for in the models. Firm controls and fixed effects identical to the baseline models are included. For brevity, we only report the estimates for the main variables of interest. Standard errors are double-clustered at user and month levels. *T*-statistics are reported in parenthesis. Symbols ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A. Stock coverage

	Dependent variable: *RET*											
	Disc. accruals		*Firm complexity*		*Idiosyncratic risk*		*Firm age*		*Analyst disagreement*		*Operating profitability volatility*	
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Sentiment	0.026	0.621**	0.261*	0.541**	-0.018	0.766***	0.722***	0.066	0.322	0.465***	-0.028	0.555***
	(0.158)	(2.074)	(1.776)	(2.139)	(-0.367)	(3.503)	(3.137)	(0.525)	(1.561)	(2.882)	(-0.245)	(2.748)
ln(# of stocks covered)	-0.125***	-0.107*	-0.076	-0.225***	0.004	-0.138***	-0.114	-0.079**	-0.134*	-0.084	-0.036	-0.087*
	(-3.177)	(-1.683)	(-1.484)	(-2.722)	(0.193)	(-2.668)	(-1.614)	(-2.291)	(-1.836)	(-1.639)	(-1.198)	(-1.722)
Sentiment. × ln(# of stocks covered)	-0.001	-0.082*	-0.038	-0.100**	0.008	-0.107***	-0.107***	-0.01	-0.044	-0.089***	0.01	-0.082***
	(-0.027)	(-1.768)	(-1.496)	(-2.425)	(0.930)	(-3.176)	(-2.817)	(-0.448)	(-1.033)	(-3.399)	(0.524)	(-2.592)
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for *Sentiment × ln(# of tweets)*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for dividing variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	704,496	1,457,457	2,311,040	430,603	667,137	1,561,659	1,600,236	1,599,786	1,168,106	2,031,916	769,067	2,241,143
Adjusted R^2	0.209	0.175	0.129	0.483	0.225	0.194	0.169	0.142	0.152	0.165	0.170	0.147
H_0: Coefficient equality for *Sentiment × ln(# of stocks covered)* across subsamples [*t*-stat]	-1.367		-1.384		-3.351***		2.147**		-0.735		-2.470**	

Panel B. Industry coverage

	Dependent variable: *RET*											
	Disc. accruals		*Firm complexity*		*Idiosyncratic risk*		*Firm age*		*Analyst disagreement*		*Operating profitability volatility*	
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Sentiment	0.029	0.597**	0.255*	0.594**	-0.029	0.754***	0.717***	0.065	0.326	0.467***	-0.04	0.531***
	(0.160)	(1.971)	(1.701)	(2.234)	(-0.581)	(3.485)	(3.170)	(0.496)	(1.515)	(2.830)	(-0.308)	(2.694)
ln(# of industries covered)	-0.152***	-0.148*	-0.092	-0.308***	0.004	-0.168***	-0.138	-0.101**	-0.165*	-0.101*	-0.041	-0.117*
	(-3.211)	(-1.958)	(-1.491)	(-2.817)	(0.185)	(-2.651)	(-1.644)	(-2.315)	(-1.918)	(-1.794)	(-1.078)	(-1.910)
Sentiment × ln(# of industries covered)	-0.002	-0.093	-0.045	-0.141**	0.013	-0.128***	-0.131***	-0.012	-0.057	-0.111***	0.016	-0.093**
	(-0.048)	(-1.580)	(-1.376)	(-2.474)	(1.251)	(-3.103)	(-2.797)	(-0.409)	(-0.993)	(-3.304)	(0.569)	(-2.435)
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for *Sentiment × ln(# of tweets)*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for dividing variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	704,496	1,457,457	2,311,040	430,603	667,137	1,561,659	1,600,236	1,599,786	1,168,106	2,031,916	769,067	2,241,143
Adjusted R^2	0.209	0.175	0.129	0.483	0.225	0.194	0.169	0.142	0.152	0.165	0.17	0.147
H_0: Coefficient equality for *Sentiment × ln(# of industries covered)* across subsamples [*t*-stat]	-1.141		-1.528		-3.323***		2.097**		-0.675		-2.272**	

4.8 Appendix: Variable Defintiions

TABLE 15
Detailed Variable Definitions

Variables	Definition	Source
Tweet Variables:		
Sentiment	Total number of bullish tweets divided by total number of tweets posted by user *i* for firm *j* in month *t*.	StockTwits
# of stocks covered	Total number of unique stocks covered by user *i* within the 6-month window ending at the beginning of month *t-1*.	StockTwits
# of industries covered	Total number of unique 3-digit SIC industries covered by user *i* within the 6-month window ending at the beginning of month *t-1*.	StockTwits
# of tweets	Total number of tweets user *i* posted for firm *j* in month *t*.	StockTwits
# of likes	Total number of likes received by user *i* from other users at the end of each 6-month window ending at the beginning of month *t*.	StockTwits
# of months each stock followed	Total number of months since firm *j* has first been tweeted about by a given user *i*.	StockTwits
Firm characteristics:		
PRC	Closing prices measured at the end of month *t-2*.	CRSP
ME (in million)	Market capitalization, calculated by absolute price times share outstanding, measured at the end of month *t-2*.	CRSP
VOL (in thousand)	Monthly stock trading volume measured at the end of month *t-2*.	
BM	Book-to-market equity ratio. Computed using book equity value in fiscal year ending in year *t-1* divided by market capitalization at the end of December in year *t-1*.	Compustat & CRSP
Dividend Yield	Dividend yield, computed as the sum of all dividends paid over the 12 months divided by share price in month *t-2*.	CRSP
RET(-2, -3) (%)	Momentum returns, computed by cumulating monthly stock returns over the period from month *t-3* to *t-2*.	CRSP
RET(-4, -6) (%)	Momentum returns, computed by cumulating monthly stock returns over the period from month *t-6* to *t-4*.	CRSP
RET(-7, -12) (%)	Momentum returns, computed by cumulating monthly stock returns over the period from month *t-12* to *t-7*.	CRSP
RET (%)	Monthly excess stock returns (in excess of the 3-month Treasury bill rate).	CRSP
ABRET (%)	Abnormal stock returns adjusted by estimating the Carhart (1997) four-factor model using the past 6 years of monthly return data ending at the beginning of the previous month. The asset pricing factors are obtained from Professor Kenneth French's website.	CRSP

INDRET (FF49) (%)	Industry-adjusted monthly stock returns by demedianing within each of 49 Fama-French industries.	CRSP
INDRET (SIC3) (%)	Industry-adjusted monthly stock returns by demedianing within each of 3-digit SIC industries.	CRSP
Analyst disagreement	Standard deviation of analyst forecasts of the earnings-per-share (EPS) long-term growth rate (LTG) for a given firm in month *t*-1 following Yu (2011). For subsample analysis, results are subsampled at 80 percentiles due to skewness.	I/B/E/S
Disc. accruals	Discretionary accruals from last fiscal year, computed as the difference between actual and "normal" accruals in a given firm and given year. Normal accrual is the model-fitted accruals, fitted using the estimated parameters from the estimation of the modified Jones model (Dechow et al., 1995) in each 3-digit SIC industry and year. Results are subsampled at 50 percentiles.	Compustat
Idiosyncratic risk	Firm's idiosyncratic risk at month *t-1*, defined as the standard deviation of the regression residuals from the estimation of Carhart (1997) four-factor model using monthly returns over the past 6 years. Results are subsampled at 50 percentiles.	CRSP
Firm complexity	A dummy for complex firms that equals one for firms who operate in more than one 2-digit SIC industry and whose aggregate sales from all reported industry segments account for more than 80% of its total sales; it equals zero for standalone firms which operate only in one industry and whose segment sales account for more than 80% of its total sales.	Compustat & CRSP
Firm age	Standard deviation of analyst forecasts of the earnings-per-share (EPS) long-term growth rate (LTG) for a given firm in month *t*-1 following Yu (2011). For subsample analysis, results are subsampled at 80 percentiles due to skewness.	CRSP
Operating profitability volatility	Standard deviation of return on asset (ROA) from last fiscal year, computed from Income before extraordinary items (IB) of current year *t* divided by Total asset (AT) from last year *t-1*. Standard deviation is calculated from 5 years window of data and require at least 3 years of ROA. Results are subsampled at 50 percentiles.	Compustat
Earnings predictability table		
SUE	Standardized Unexpected Earning surprises, calculated from the difference between this and same quarter from last fiscal year's EPS scaled by last quarter's stock price.	I/B/E/S
# of analyst	Number of analysts covering firm i one month prior to the earning-announcement's month.	I/B/E/S
Fiscal year end dummy	A dummy variable that equals one when the earning announcement is made at the end of the fiscal year, and zero otherwise.	Compustat & I/B/E/S
Last quarter loss dummy	A dummy variable that equals one when a firm incur losses during the previous quarter, and zero otherwise.	I/B/E/S

4.9 Summary for Chapter 4

This chapter studies the implications of portfolio complexity for the quality of stock analysis among individual investors. Individuals who prefer following and analysing many stocks, and/or stocks from many industries simultaneously have more complex portfolios; whereas those following only a few are less complex.

Researching stocks requires significant time and effort and typically involves gathering, analysing and interpreting financial information. On the one hand, investors following multiple stocks likely have superior market-wide knowledge. On the other, those specializing in only a few have a comparative advantage in understanding the followed companies in greater depth. Nevertheless, since human beings have limited cognitive resources to spread over tasks, attention spent on one task necessarily reduces the attention available for others (Kahneman, 1973), suggesting that the quality of stock analysis may deteriorate as the number of stocks users follow increases.

To test this hypothesis, all tweets were downloaded from StockTwits, one of the oldest and widely-used social-network platforms in stock trading. Users post micro-blog messages (tweets) about companies to share trading ideas. Using machine-learning techniques to classify social-media tweets into {*Bullish, Neutral, Bearish*} sentiment category, the number of unique stocks and industries was tracked with users who tweeted. A Sentiment value was then created by aggregating all *Bullish* and *Bearish* tweets for each user, firm, and month, removing *Neutral* tweets because they conveyed information unrelated to the stock discussion. The result showed that social-media sentiment positively and significantly predicted future stock returns, controlling for a wide range of firm characteristics and user and month fixed effects. Importantly, it was found that the positive predictability of social-media sentiment decreases with the number of stocks and industries followed. Such results are robust to a battery of robustness checks and controls for users' reputation and prior experience with the followed stocks.

To test for the limited-attention hypothesis, my results continued to hold when excluding users who declare themselves technical and short-term traders, who did not formulate trading ideas based on company fundamental news. Next, consistent with users experiencing limited attention and reduced research effort on company fundamental positions, they predicted positively and significantly a company's earnings, and such predictability decreased with their expanding stock and industry coverage. Furthermore, the loss in predictability due to increasing

stock and industry coverage over future stock returns was only significant among firms that were harder to analyse due to more complex organizational structures and lower informational transparency.

Finally, a practical trading strategy was designed to exploit user characteristics by dividing users into low and high stock-coverage groups and computing the low-minus-high average sentiment for each stock. The strategy longed a portfolio of stocks with positive and zero net sentiment and shorted a portfolio of stocks with negative net sentiment which generated a significant trading alpha of 26.4 basis points per month.

This study contributes to the literature in several ways. The findings add to the growing body of literature examining the implications of investor sentiment for asset prices (see, e.g., Baker and Wrugler, 2006; Baker et al., 2012; Da et al., 2015). Numerous studies estimate social-media sentiment at the firm level. For instance, measuring sentiment using Twitter data, Bartov et al. (2018) examined its link with future earnings and stock returns. Cookson and Niessner (2019) investigated how disagreement between StockTwits users drives trading volume. I also contributed to the experimental finance literature, particularly in Bossaerts et al. (2020). Their paper studies individual computational complexity where assets uncertainty is the consequence of not knowing which information is pertinent rather than not having the information in the first place. My findings complement the above studies.

5. Social-Media Disagreement and the concavity of SML line

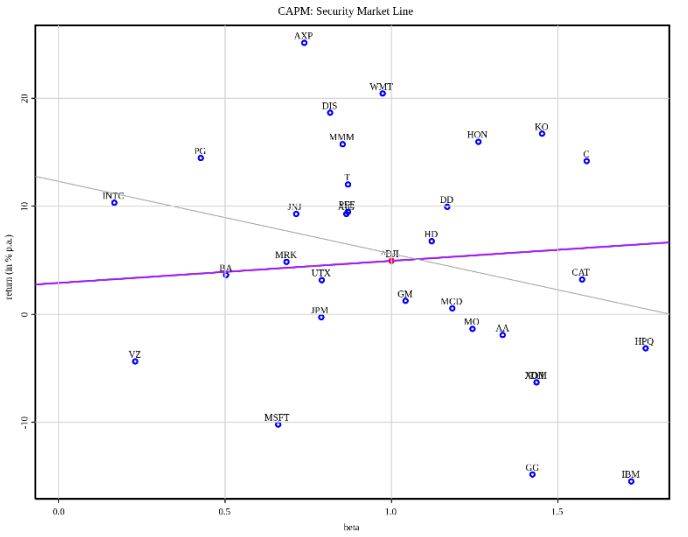
5.1 Introduction

Prior empirical research has agreed on a well-known fact that assets with high risk do not necessarily deliver a higher return, contradicting the prediction from the traditional asset pricing theory. That is to say, investors who invest in high-risk stocks do not receive the required compensation for their risks taken. This high-risk, low return puzzle dates back to Black (1972), Black, Jensen, and Scholes (1972), where they have shown high-risk stocks, as captured by high CAPM beta, have underperformed low-risk in the last 30 years. In other words, the empirical Securities Market Line (SML), as shown by the tradeoff between expected return and beta, is a downward sloping against CAPM's prediction. Figure 9 provides the relationships between expected return and beta risk graphically. The purple (gray) line shows the CAPM predicted SML line (empirically fitted SML line) using return data from the Dow Jones Industrial Average over three years of monthly data. Observe that the purple line indicates the CAPM predicted SML line with a positive and linear relationship between expected return and risk. In contrast, the empirically gray SML line is a downward sloping.

Figure 9

CAPM's Securities Market Line and Empirically Fitted SML Line

Fig. 9 shows the CAPM's Securities Market Line (purple) and the empirically fitted SML line (gray). The market portfolio uses Dow Jones Industrial Average over three years for monthly data. (Courtesy of Wikipedia)



Baker, Bradley, and Wurgler (2011) were the first to conceptualize such anomaly in monetary terms. They showed that cumulative performance of stocks starting from January 1968 declines with beta. If one invests a dollar in a value-weighted portfolio of the lowest quintile of beta stocks, he would have yielded $96.21 ($15.35 in real terms) in comparison to $26.39 ($4.21 in real terms) by investing in the highest quintile beta stocks. Frazzini and Pedersen (2014) found that one can exploit this anomaly by forming a "Market-Neutral Bet Against Beta" factor trading strategy. Such a strategy exploits the mispricing by constructing a portfolio that holds low-beta stocks, leveraged to a beta of one, and shorts high-beta stocks, de-leveraged to a beta of one. This BAB strategy achieves a zero beta portfolio in US stocks that hold $1.4 of low-beta stocks and use the proceeds to short-sell $0.7 of high-beta stocks, together with using offsetting position in the risk-free asset to make it self-financing. This strategy generates excess profits consistent with other factors documented with Fama and French (1993) and Carhart (1994) size, growth, and momentum.

Given all the profitable strategies, and more importantly, the anomaly does not disappear, academics have yet to explain why the SML line is flat. Indeed, Black (1972) was the first to re-examine the original CAPM model and modify the central assumption theoretically. He relaxed the assumption of restricted borrowing or lending opportunities at a risk-free rate. There is a risk-free asset in the model, but one would incur some borrowing or lending cost through the risk-free rate. The investor would come up with investing in a zero-beta portfolio that could replace the risk-free asset. This portfolio has a beta of zero in a sense there is no co-movement with the market. In reality, one could form a zero beta portfolio by, for example, buying a commodity with a beta of -1, using the proceeds to short-sell a stock market portfolio with beta equals 1. The model is as follows:

$$E(R_i) = E(R_z) + \beta\,[E(R_m) - E(R_z)] \qquad (4)$$

R_z denote the return from the zero-beta portfolio and represents the intercept in the y-axis following an expected return $E(R_i)$ versus β graph. R_i and R_m are assets i and market returns. Risk-free asset return by construction is non-negative $R_f \geq 0$ and is the guaranteed return one should receive. The zero-beta portfolio would always receive a higher expected return than the risk-free rate. I.e., $E(R_z) > R_f$. Hence, when a zero-beta portfolio replaces the risk-free asset, one would get a flatter theoretical SML line closer to empirically revealed. I.e., $E(R_m - R_z) < E(R_m - R_f)$.

Frazzini and Pederson (2014) proposed a model that borrows the idea from Black (1972) of restricted borrowing, whereby it suggests individuals have margin constraints. The model assumes that investors generally opt for the highest Sharpe ratio (Excess return per unit of risk) investment and can be represented by three types of agents. Margin constrained investor 1 cannot leverage and therefore weigh more on high-beta assets in their portfolio, making high-beta stocks overpriced and lower expected return. Investor 2 can leverage but faces margin calls. Margin unconstrained investor 3 can buy and sell any assets freely. This model is similar to Black (1972) model, except the SML line's slope also depends on the tightness of the funding constraint on average across agents. Where the funding constraint can understand as how much an agent can borrow money to leverage.

Frazzini and Pederson (2014) asset pricing equation with funding constraint is as follows:

$$E_t(R^s_{t+1}) = R_f + \varepsilon_t + \beta^s_t \lambda_t$$

$$\lambda_t = E_t(R^M_{t+1}) - R_f - \varepsilon_t \qquad (5)$$

Where s is securities $s = 1, \ldots, s$, ε_t is the average Lagrange multiplier which measures the tightness of funding constraint.

A higher value indicates a binding investor's utility function to the funding constraint. β^s_t is the sensitivity of asset s in relation to the future market return R^M_{t+1}. λ_t is the market risk premium and is represented by $\lambda_t = E_t(R^M_{t+1}) - R_f - \varepsilon_t$. One can see that ε_t is in λ_t and if funding constraint tightens $\varepsilon_t > 0$, the first effect is market risk premium λ_t gets smaller. Secondly, the intercept, i.e., $R_f + \varepsilon_t$ increases, and the SML line becomes flatter. In other words, when the slope is smaller, a constrained investor can only hope for a non-leveraged return and be willing to accept less compensation for higher risk. Armed with such theory, they developed a new Bet Against Beta (BAB) factor investing strategy.

Equation (6) below shows the BAB portfolio return equation:

$$R^{BAB}_{t+1} = 1/\beta^L_t \, [R^L_{t+1} - R_f] - 1/\beta^H_t \, [R^H_{t+1} - R_f] \qquad (6)$$

L and H represent low and high beta portfolios of stocks, and $\beta^H_t > \beta^L_t$. R^L_{t+1}, R^H_{t+1} are the next period return for the low and high portfolios, respectively. The portfolio is formed by ranking each stock's beta into two portfolios, i.e., low and high beta portfolios, using the median of stocks each month and weighted by beta.

Empirically, they collected data spanning multiple asset classes from January 1926

to March 2012 and tested several model predictions. Consistent with Black (1972), the prediction where alpha declines monotonically as beta increases, implying that the Sharpe ratio is also declining. In the second test, they test that when the economy is experiencing tightened funding constraints, contemporaneous BAB factor return will be negative because the required future BAB return is increased to reflect the funding risk investors need to bear. Using TED spread as a proxy for funding constraint when BAB return is regressed on lagged, and contemporaneous TED spread, both coefficients exhibit negative relationships, inconsistent with their theoretical model predicted. Instead, they argued this is due to numerous factors that the model has not taken into account.

Hong and Sraer (2016) provided a new perspective by dropping more CAPM assumptions and including behavioural components and individual short-selling constraints. Such relaxation allows investors to have heterogeneous expectations about the asset's end-of-period value. The assumption of heterogeneous investor expectations has been validated by significant empirical evidence showing time-varying disagreement between industry analysts and individual investors who hold different expectations on macroeconomic variables. For example, market earnings, inflation, industrial production growth from Cukierman and Wachtel (1979), Kandel and Pearson (1995), Mankiw, Reis, and Wolfers (2004), and Lamont (2002). More recently, Diether, Malloy, and Scherbina (2002) found that individual stocks also exhibit disagreement among stock analysts as they provide guided EPS estimates, see chapter 2 literature review.

With disagreement and short-selling constraints, Hong and Sraer's model predicts an inverted-U shape SML line that matches the empirical data. In contrast to Frazzini and Pedersen (2014), the slope of SML depends on the tightness of funding constraint; their model predicts that the high concaveness (or downward sloping) is due to disagreement on the common factor of firm's cash flow. Investors disagree on average the common factor of a stock, and such disagreement is naturally higher in high-beta than low-beta risk stocks. To illustrate this idea, suppose that a utility and a start-up company exist in the economy, the utility company tends to have lower market risk and stable cash flow because every household requires electricity, making investors less likely to disagree about their cash flow. In contrast, investors generally disagree more with a start-up company because they carry more uncertainty to their cash flow and have higher market risk.

A macro-scale disagreement can be proxied by aggregating these stocks-level disagreement weights by their respective beta; for example, high-beta stocks amplify a macroeconomy disagreement due to them being more uncertain and disagreeing more generally.

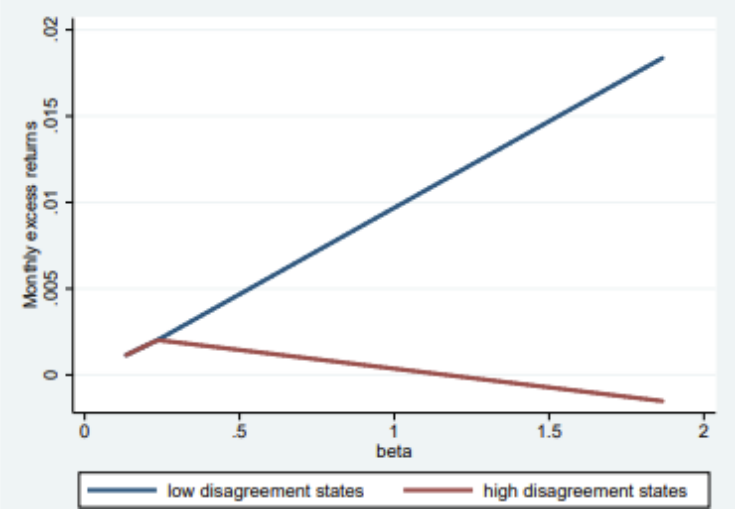
Having established the positive relationships between beta and disagreement, the model by Hong and Sraer (2016) requires one last puzzle piece: short-sell constraint. When the economy is at a low aggregate disagreement state, all stocks do not generate high enough disagreement irrespective of their beta. Short-sell constraints do not bind because investors do not see the reason to short sell, retaining the traditional CAPM's SML high risk, high return characteristics.

In contrast, less is affected to low-beta stocks during the high aggregate disagreement state because of their low stock-level disagreement. Optimists overwhelm pessimist investors, and their short-sale constraints do not bind. For increasingly high-beta stocks, as the stock disagreement increases, both optimist and pessimist investors reach a point where short-sell constraint binds. Pessimists who wish to sell stocks are sidelined and overprice the high-beta stocks, bending the SML line to a concave shape. Figure 10 below illustrates the idea of the SML line under low and high aggregate disagreement states. Notice the kinked line in the high aggregate disagreement state (red line).

Figure 10

Securities Market Line Under Low and High Aggregate Disagreement States

Fig. 10 shows Hong and Sraer (2016)'s model simulation on the securities market line under low aggregate disagreement (blue line) and high aggregate disagreement (red line).



De Giorgi, Post, and Yalcin (2020) took a different approach to explain the concavity of the SML line. In their model, unlike the argument of disagreement from investors, the driving force of the SML line's concavity is from a pool of investors with varying risk aversion holding different sets of risky securities and some with ever binding short-selling constraints in their portfolios. In their model, every investor's portfolio is generally exposed and linearly related to the same common factor, i.e., systematic risk. Without knowing the composition of each investor's portfolio, for example, risk aversion parameters, portfolio wealth, etc. There will be no case for testing the general equilibrium empirically. However, given individual investors' exposure to systematic risk, one can back out, aggregate, and approximate a concave, piecewise linear relationship between expected return and beta. To empirically test the SML concavity with their theory, they share the same Fama and MacBeth (1973) style regression and reports result similar to Hong and Sraer (2016) robust findings after controlling other stock characteristics. However, the beta-squared coefficient is insignificant in some sample periods, suggesting the empirical concave SML line is time-varying.

Having understood the background in this strand of literature, I followed Hong and Sraer's (2016) empirical testing procedure using my social-media dataset to obtain social-media disagreement and tested whether such explains the concavity of the SML line. Their original contribution focused on aggregate disagreement using I/B/E/S analyst forecast dispersion of EPS estimates in Long-Term-Growth (LTG) and argued that such measure featured prominently in valuation models. My result indicates that aggregate social-media disagreement significantly explains the curvature of SML line in future 9-, 12-, 18-month excess return and results are stronger as return interval is higher.

To the extent of using which proxy for disagreement fares better, I argued that using social-media disagreement is a more appropriate measure because, firstly, the primary users of professional forecasters are institutional investors. An individual investor does not have access to such information. Secondly, as suggested in section 2.3, institutional investors, i.e., Hedge Fund, also use social media to trade ideas, representing the balanced thoughts between individual and institutional investors. Third, the most important part is that individual investors are usually the primary group that faces short-sell constraints, whereas institutional investors are more flexible in this regard. Social-media disagreement has incorporated all these pieces of information and reflects in a more "down to earth" approach.

5.2 Measuring social-media disagreement

In this section, the details are outlined on how to calculate the firm-level social-media disagreement. Tweets are firstly classified using machine learning algorithm as suggested in section 3 into categorical variable {Bullish, Neutral, Bearish} and discard any Neutral tweets. Since prior disagreement measures are continuous, disagreement can be reliably measured using second moments, i.e., standard deviation. I followed Antweiler and Frank (2004) by turning categorial to continuous variable and calculate the disagreement measure. This is accomplished by firstly calculating the *Sentiment*:

$$Sentiment_{i,t} = (\#\ of\ bullish\ tweets)_{i,t} - (\#\ of\ bearish\ tweets)_{i,t} / (\#\ of\ bullish\ tweets)_{i,t} + (\#\ of\ bearish\ tweets)_{i,t} \quad (7)$$

Where *i* denotes firm *i* at month *t*, this measure is bound between -1 and +1, with the extremes representing bearish and bullish, respectively. The treatment for periods with no tweets are posted for both *# of bullish tweets and # of bearish tweets* as zero and *Sentiment* of that month to be zero, assumed that no one has anything to say and is neutral. The variance or the disagreement of *Sentiment* for firm *i* at month *t* corresponding to equation (7) can be calculated as:

$$Disagree_{i,t} = \sqrt{1 - Sentiment^2_{i,t}} \in [0,1] \quad (8)$$

Disagree is bounded between 0 and 1 and is higher when disagreement is high.[32] To see the properties of *Disagree*, suppose we have three bullish tweets. *Sentiment* will be 1, and *Disagree* is 0, meaning that social-media investors all agree. Suppose there are two tweets for bullish and bearish, $Sentiment = 0/4 = 0$ and $Disagree = \sqrt{1 - 0^2} = 1$, representing this firm is in high disagreement.[33]

[32] I also tried different calculation of *Disagree*, for example $Disagree = 1 - |Sentiment|$ and following Li and Li (2014) by using Weighted Negative Herfindahl Index (WNHI) as $Disagree = -\Sigma w_k p^2_k$ where $k \in \{Bullish, Bearish\}$, w_k is the weight or the intensity of k category which in my case it will always assume to be 1, and p is the percentage number of tweet posted for category k. I find that results are quantitatively similar between all measures.

[33] In reality this is down to subjective judgement and I find no difference to the regression result by hard coding no tweets message firms' *Disagree* to 0 (1) as latent agreement (disagreement) and reveal that only sample size change.

5.3 Data and Variables

The U.S. return data comes from CRSP tape and social-media disagreement, as outlined in section 5.2. The sample period is limited to December 2010 to December 2017 due to the StockTwits data availability. For each month, the penny stocks is excluded with a share price below $5 and microcaps where it is defined as stocks in the bottom two deciles of the monthly market capitalization distribution using NYSE breakpoints. When calculating a stock's beta, value-weighted market return (provided by Professor Ken French's website) in excess of the US Treasury bill rate is used.

5.3.1 Constructing beta-sorted portfolios

To construct beta-sorted portfolios, Hong and Sraer (2016) is followed by estimating the beta of each stock in the cross-section using the past 12 months of daily return. Regressors contain contemporaneous and five lags of excess market return following Dimson (1979) to reduce the effect of small stocks illiquidity. The beta of the stock is then the sum of the six coefficients in the regression. Next, stocks are sorted into 20 beta portfolios based on these pre-ranking betas by using only stocks in NYSE to define the thresholds, daily returns on these portfolios, both equal- and value-weighted returns are calculated. The portfolio's post-ranking beta is estimated by regressing the excess portfolio return on contemporaneous excess market return and five lags, totalling the six OLS coefficients using the full sample period following Fama and French (1992).

Table 16 shows the characteristics of the 20 beta-sorted portfolios. One can see the β is monotonically increasing from 0.21 in portfolio 1 to 1.79 in portfolio 20. The forward 1-month return does not show a clear relationship between beta and return; however, the forward 12-month return does show a clear inverted U-shape of return and beta. A return of 3.32% in portfolio 1 is observed to the highest return of 14.33 at portfolio 8 and a decreasing return of 7.94% in portfolio 20, suggesting a nonlinear relationship between return and portfolio beta. *Median volatility* is the median pre-ranking volatility of the stocks in the portfolio. An increasing linear trend across all portfolios can be seen from portfolio 1 of 0.69 to portfolio 20 of 2.88, consistent with the idea that higher beta stocks also have higher volatility. *% Market Capitalization* also shows a concave pattern as portfolio beta increases. Stocks with the highest market capitalization concentrate at a portfolio beta of 1 with around 6% of market

capitalization. In comparison, the most extreme portfolios of 1 (20) are only 1.61% (2.73%), respectively.

% short, defined as the percentage of short interest, is calculated by dividing the short interest by the share outstanding in the sample. Again it is an increasing linear trend across portfolios from 2.02% in portfolio 1 to 6.34% in the highest beta portfolio 20. Short interest represents stocks that are shorted in the market but have not been covered. Figlewski (1981) uses the short interest to proxy for short sale demand and argues firms with high short interest are more difficult to short. The table shows a direct, robust, and consistent phenomenon where higher beta stocks are harder to short, hence the high short interest. Lastly, consistent with Hong and Sraer (2016) model prediction, social-media disagreement is monotonically increasing as the portfolio exhibits higher beta, shown by *Stock Disp.* from 0.508 in portfolio 1 to 0.636 in portfolio 20.

TABLE 16
Summary Statistics for 20 beta-sorted Portfolios

Sample period: 12/2010 to 12/2017. Sample: CRSP stocks file excluding penny stocks (price < $5) and microcap (stocks in the bottom two deciles of the monthly size distribution using NYSE break points). At the beginning of each calendar month, stocks are ranked in ascending order on the basis of their estimated beta at the end of the previous month. Preformation betas are estimated with a market model using daily returns over the past 12 months and five lags of market return. The ranked stocks are assigned 1 to 20 value-weighted portfolio using NYSE as break points. The table reports the full sample β of each of these 20 portfolios, computed using a similar risk model. *Median Volatility* is the median pre-ranking volatility of stocks in the portfolio. $R^{(1)}_{i,t}$ is the return of portfolio from *t* to *t+1*, $R^{(12)}_{i,t}$ is the return of portfolio from *t* to *t+11*. *Stock Disp.* is the average of Social-Media disagreement. *% Market Cap.* Is the average ratio of market capitalization of stocks in the portfolio divided by the total market capitalization of stocks in the sample. *% Short* is the percentage of short interest in the sample, calculated by short interest divided by share outstanding. *N stocks* is the number of stocks on average in each portfolio.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
β	0.21	0.51	0.57	0.65	0.72	0.79	0.84	0.9	0.98	1	1.01	1.1	1.13	1.19	1.22	1.3	1.32	1.48	1.5	1.79
Median Volatility	0.69	1.04	1.13	1.16	1.2	1.2	1.22	1.23	1.23	1.23	1.28	1.31	1.37	1.48	1.6	1.7	1.83	2	2.22	2.88
$R^{(1)}_{I,t}$	0.02	1.01	0.94	0.96	1.24	0.74	1.14	0.85	1.27	0.84	0.93	0.94	0.57	0.77	0.81	0.2	0.46	0.76	0.74	0.97
$R^{(12)}_{I,t}$	3.32	10.69	11.69	11.36	13.63	13.38	12.57	14.33	13.12	12.44	11.59	12.03	12.33	10.73	11.68	10.74	9.76	8.17	7.71	7.94
Stock Disp.	0.508	0.511	0.510	0.522	0.515	0.506	0.514	0.519	0.521	0.518	0.525	0.528	0.535	0.549	0.564	0.566	0.564	0.570	0.590	0.636
% Market Cap.	1.61	5.02	5.92	6.26	6.03	6.11	6.53	7.44	6.87	6.91	6.51	6.5	6.6	6.1	5.81	5.65	5.36	5.44	3.72	2.73
% Short	2.02	2.44	2.29	2.45	2.59	2.69	2.81	3.02	3.07	3.14	3.15	3.21	3.24	3.31	3.5	3.65	3.9	4.31	4.74	6.34
N stocks	95	94	105	117	116	123	127	134	139	141	147	148	146	147	147	146	148	152	150	190

5.3.2 Constructing Aggregate Social-Media Disagreement

Aggregate social-media disagreement is a bottom-up measure constructed in spirit to Yu (2011). The idea is that using stock-level social-media disagreement, one can aggregate to market level and proxies for the macroeconomic, market-wide disagreement. To construct this measure, the stock-level social-media disagreement *Disagree* is used and all stocks in the sample are aggregated using pre-ranking beta $\beta_{i,t-1}$ for each stock as weight.

$$Agg.\ Disagree_t = \Sigma_i \left(\beta_{i,t-1}\ Disagree_{i,t} / \Sigma_i\ \beta_{i,t-1}\right) \qquad (9)$$

Where *i* represents stocks and month at *t*, the *Agg. Disagree* at time *t* is calculated by summing and multiplying all stock's pre-ranking beta calculated from last month in the sample and their respective disagreement value, Disagree, is divided by the pre-ranking beta.

Another version of *Agg. Disagree* is also tested, which considers the beta mean-reverting properties over time towards the market of 1. In this case, equal weights have been given to the market beta of 1 and stock beta. The two measures have a minimal difference in the subsequent regression analysis because we essentially multiply and add a constant.

$$Agg.\ Disagree_t\ (Compressed) = \Sigma_i \left(\ (0.5\ \beta_{i,t-1} + 0.5)\ Disagree_{i,t} / \Sigma_i\ (0.5\ \beta_{i,t-1} + 0.5)\ \right) \qquad (10)$$

Following Hong and Sraer (2016), using beta as weight is justified by the fact that their model presents where overall disagreement on stock's dividend d_i contains two components i.e., $d_i = d + b_i z + \varepsilon_i$. The first is the aggregate factor z (i.e., the economy), and the second is the idiosyncratic factor ε_i. Suppose stock *i*'s beta is at 0, i.e., there is no sensitivity to market risk at all; most of the risk comes from idiosyncratic disagreement. When beta is getting larger and more sensitive to the market risk, investors generally disagree more about the market but less on the idiosyncratic component.

Figure 11
Time Series of Standardized Aggregate Social-Media Disagreement

This figure shows the aggregate social-media disagreement from 12/2010 to 12/2017. For each month, I calculate for each stock's social-media disagreement. I then estimate the stock-level beta from the previous month using a market model with daily returns over the past 12 months and five lags for market returns. The aggregate social-media disagreement is the monthly beta-weighted stock level disagreement using beta from the previous month and is standardized for easier interpretation.

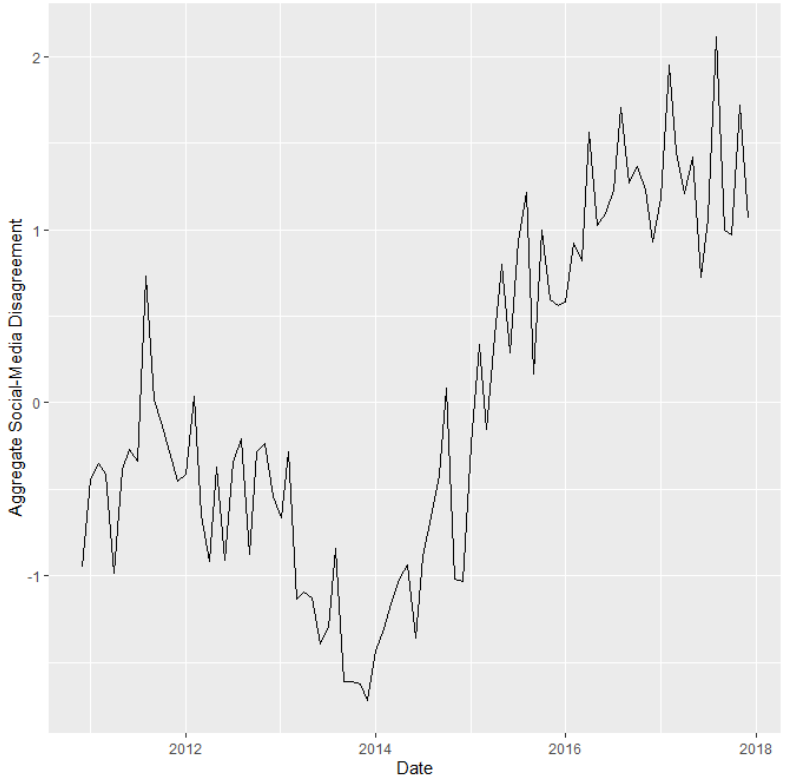


Figure 11 above shows the time series of standardized aggregate social-media disagreement from 12/2010 to 12/2017. It mimics Figure 5 in section 4.2 in that social-media users have a divergence of bull and bear tweets starting from the year 2015. For example, one can see December 2014 in Figure 11 has the lowest aggregate social-media disagreement, consistent with the bull market during the end of 2013.[34] A steady increase in disagreement is observed going up. It settles at around 2016, most likely due to the US presidential election, where political uncertainty plays a role in the stock market. Table 17 below shows the summary statistics on the time-series variable used in this analysis. All aggregate disagreement variables are standardized for easier interpretation.

[34] https://www.cnbc.com/2013/12/31/us-stocks.html

TABLE 17
Summary Statistics for time-series variables

Sample period: 12/2010 to 12/2017. Sample: CRSP stocks file excluding penny stocks (price < $5) and microcap (stocks in the bottom two deciles of the monthly size distribution using NYSE break points). At the beginning of each calendar month, stocks are ranked in ascending order on the basis of their estimated beta at the end of the previous month. Preformation betas are estimated with a market model using daily returns over the past 12 months and five lags of market return. The ranked stocks are assigned 1 to 20 value-weighted portfolio using NYSE as break points. The table reports the summary statistics for time-series variables. *Agg. Disagree* is the beta weighted stock level social-media disagreement using beta from previous month. *Agg. Disagree (Compressed)* uses $0.5\,\beta + 0.5$ as the weight for stock level social-media disagreement. $R^{(12)}_{m,t}$, $SMB^{(12)}_t$, $HML^{(12)}_t$, $UMD^{(12)}_t$ are the 12-month monthly rolling return on the market, SMB, HML, UMD portfolios from Ken French's website and are represented as percentage. D/P and E/P are the Dividend-Price ratio and Earning-Price ratio respectively, obtained from Robert Shiller's website. Inflation is the yearly inflation rate. TED Spread is the difference between three-month Treasury bill and three-month LIBOR based in US dollars. *VIX* is the CBOE volatility index, Industrial Production Index obtained from FRED. *CAY* is consumption wealth ratio obtained from Martin Lettau website. *Short-term interest* obtained from OECD website. *Default Spread* is the Baa corporate bond minus 10-year maturity treasury bond and lastly *Term Spread* is the 10-year treasury minus 3-month treasury.

	N	Mean	St. Dev.	Min	p(25)	p(75)	Max
Agg. Disagree	72	0	1	-1.688	-0.798	0.861	2.134
Agg. Disagree (Compressed)	72	0	1	-1.709	-0.805	0.865	2.114
$R^{(12)}_{m,t}$	72	13.892	9.856	-8.372	4.378	21.352	35.166
$SMB^{(12)}_t$	72	-0.897	4.963	-11.76	-5.096	3.122	8.74
$HML^{(12)}_t$	72	0.639	7.601	-13.76	-4.627	5.035	20.716
$UMD^{(12)}_t$	72	2.174	9.102	-19.5	-2.054	6.604	22.132
D/P	72	2.025	0.114	1.756	1.949	2.112	2.296
E/P	72	0.055	0.009	0.041	0.046	0.062	0.074
Inflation	72	0.137	0.333	-0.567	-0.143	0.38	0.975
TED spread	72	0.285	0.106	0.154	0.213	0.357	0.561
VIX	72	17.349	5.267	11.4	13.87	18.415	42.96
Industrial Production	72	101.64	2.881	95.515	99.904	103.73	106.66
CAY	72	-0.011	0.01	-0.027	-0.018	-0.005	0.007
Short-term interest	72	0.283	0.173	0.11	0.14	0.305	0.75
Default Spread	72	2.809	0.339	2.19	2.57	3.12	3.56
Term Spread	72	2.159	0.552	1.2	1.708	2.552	3.44

Table 17 shows the summary statistics for all the time-series variables used in this analysis. *Agg. Disagree* is the standardized aggregate disagreement aggregated from stock-level social-media disagreement and scaled with stock's beta. *Agg. Disagree (Compressed)* measured the same as *Agg. Disagree* with the exception that it adjusts for the property of all stock's beta shrinks towards beta of 1 (the market beta). Both measures show similar characteristics with a standard deviation of 1 which do not appear skewed, as shown by the 25 and 75 percentile. $R^{(12)}_{m,t}$, $SMB^{(12)}_t$, $HML^{(12)}_t$, $UMD^{(12)}_t$ are the 12-month monthly rolling return on the market, SMB,

HML, UMD portfolios from Ken French's website which are represented as a percentage. On average, investing in the market portfolio and holding for 1 year has an average of 13.892% return with the worse (highest) return of -8.372% (35.166%) for the sample period between 12/2010 to 12/2017. Interestingly, the *SMB* factor shows a negative average return of -0.897% per month, which is against the finding in early years by Fama and French (1992) that small firms outperform larger firms. *HML* appears to be consistent with the literature for this sample period that value firms outperform growth firms, with an average return of 0.639% per month. *UMD* is the momentum factor that calculates from taking the difference between two equal-weighted portfolios from the top 50% highest and lowest performing stocks in the portfolios. Mean shows a return of 2.174% per month, accompanied by the highest return of 22.132% and the lowest return of -19.5% per month.

D/P and *E/P* are the market portfolio's Dividend-price and Earning-price ratios, respectively. It shows the macroeconomic performance on the market as a whole on how much market willingness to provide dividends and the earnings generated. All show relatively stable statistics as demonstrated by the low standard deviation for both time series. *Inflation* is the yearly inflation rate; during the sample period, the economy experienced deflation of -0.567% for the whole year, while maximum inflation of 0.975% is shown. *TED-Spread* shows the difference between the three-month Treasury bill and three-month LIBOR based on US dollars. A higher *TED-Spread* means interbank lending is costly and could be inferred as an increase in credit risk. The summary statistics show a mean of 0.285 with a standard deviation of 0.106.

VIX is the CBOE volatility index that uses the options and futures to proxy for the future volatility of the securities. *Industrial Production* is an economic indicator in the US that measures real output for mining, electric, and gas production and its proxies for macroeconomic output like GDP. *CAY* is the consumption-wealth ratio that measures US household wealth, and a higher ratio means more wealth, Lettau and Ludvigson (2001) suggest it correlates positively with future stock returns. All series do not exhibit any outlier according to their summary statistics.

Lastly, the *Short-term interest* obtained from OECD website is the money market rate. *Default Spread* is the Baa corporate bond minus 10-year maturity treasury bond, and *Term Spread* is the 10-year treasury minus 3-month treasury. All these series represent the healthiness of an economy. For example, a higher *Default Spread* can reflect the economic environment where the

treasury bond interest is higher than the Baa corporate bond that signals investors fleeing for the safer asset, i.e., treasury bond. Again, all series do not show any anomalies. *Short-term interest* has a lower percentage return than others because of its short duration of borrowing. *Term Spread* measures the spread between different maturity dates for the safest asset, which should, in theory, be between the other twos in terms of percentage return.

5.4 Determinant of aggregate social-media disagreement?

It is a well-known fact that macro-economic factor is the main driver of investor behaviour in the stock market through economic growth. Sadorsky (2003) suggested that the business cycle affects the current value of firms. For example, more profit and earnings are earned when an economy grows, leading to better firm expectations. A higher inflation signals interest rate would rise in the future, reducing the amount of money invested in the stock market by moving money back to saving. Rising unemployment also have an adverse effect on the stock market as firms' prospect are more uncertain. However, it is unclear how macro-economic factors affect investors' aggregate social-media disagreement. It is hypothesized that when the macro-economic factors are more uncertain or becoming pessimistic, the aggregate social-media disagreement would increase. This section explores this line of inquiry by examining several macro-economic variables to predict aggregate social-media disagreement with the following regression model:

$$Agg.\ Disagree_t\ [Agg.\ Disagree_t\ (Compressed)] = \alpha + \beta_1 DP_{t-1} + \beta_2\ EP_{t-1} + \beta_3\ Inflation_{t-1} + \beta_4\ Ted\ Spread_{t-1} + \beta_5\ VIX_{t-1} + \beta_6\ Industrial\ Production_{t-1} + \beta_7\ CAY_{t-1} + \beta_8\ Short\ Term\ Interest_{t-1} + \beta_9\ Term\ Spread_{t-1} + \beta_{10}\ Default\ Spread_{t-1} + \varepsilon_t \quad (11)$$

Where *t* denotes month, *Agg. Disagree* is the aggregate social-media sentiment or Agg. Disagree (Compressed) is the compressed aggregate social-media sentiment at month *t* depending on specification is regressed on dividend-price ratio *DP (t-1)*, earning-price ratio *EP (t-1)*, rolling year-over-year inflation rate *Inflation (t-1)*, *Ted Spread (t-1)* is the difference between three-month Treasury bill and three-month LIBOR based in US dollars. *VIX (t-1)* is the CBOE volatility index, *Industrial Production (t-1)* measures the real total output in US, *CAY (t-1)* is the consumption wealth ratio, *Short Term Interest (t-1)* is the 3-month money market rate, *Default Spread (t-1)* is the spread between Baa corporate bond minus 10-year maturity bond and lastly

Term Spread (t-1) is the 10-year treasury minus 3-month treasury. All explanatory variables are lagged by one month except *CAY*, which is lagged by one quarter due to its update frequency. All variables are standardized except those represented as a percentage. Standard errors are adjusted for serial correlation using Newey West (1987) with 12 lags.

TABLE 18
Determinant Of Aggregate Social-Media Disagreement

Sample period: 12/2010 to 12/2017. The table reports the regression result of determinant of aggregate social-media disagreement. The time-series regression has the following specification:

$$Agg.\ Disagree_t\ [Agg.\ Disagree_t\ (Compressed)] = \alpha + \beta_1 DP_{t-1} + \beta_2\ EP_{t-1} + \beta_3\ Inflation_{t-1} + \beta_4\ Ted\ Spread_{t-1} + \beta_5\ VIX_{t-1} + \beta_6\ Industrial\ Production_{t-1} + \beta_7\ CAY_{q-1} + \beta_8\ Short\ Term\ Interest_{t-1} + \beta_9\ Term\ Spread_{t-1} + \beta_{10}\ Default\ Spread_{t-1} + \varepsilon_t$$

Agg. Disagree is the beta weighted stock level social-media disagreement using beta from previous month. *Agg. Disagree (Compressed)* uses $0.5\ \beta + 0.5$ as the weight for stock level social-media disagreement. *D/P* and *E/P* are the Dividend-Price ratio and Earning-Price ratio respectively, obtained from Robert Shiller's website. Inflation is the year-over-year inflation rate. *TED Spread* is the difference between three-month LIBOR based in US dollars and three-month Treasury bill. *VIX* is the CBOE volatility index, Industrial Production Index obtained from FRED. *CAY* is consumption wealth ratio obtained from Martin Lettau website. *Short-term interest* obtained from OECD website. *Default Spread* is the Baa corporate bond minus 10-year maturity treasury bond and lastly *Term Spread* is the 10-year treasury minus 3-month treasury. Standard errors are adjusted using Newey West (1987) to correct for serial-correlation using 12 lags. t-statistics are in parenthesis and ***, **, * denotes 1%, 5%, 10% significance respectively.

	Agg. Disagree (t)			*Agg. Disagree [Compressed] (t)*		
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)
DP (t-1)	0.252*	0.013	-0.325***	0.252*	0.006	-0.331***
	(1.801)	(0.250)	(-2.707)	(1.718)	(0.115)	(-2.819)
EP (t-1)	-0.520**	-0.446***	-0.343**	-0.500**	-0.430***	-0.334**
	(-2.495)	(-4.059)	(-2.398)	(-2.350)	(-3.883)	(-2.424)
Inflation (t-1)		-0.292**	-0.294*		-0.317***	-0.322**
		(-2.522)	(-1.943)		(-2.816)	(-2.170)
Ted Spread (t-1)		0.413***	0.247		0.420***	0.266
		(4.528)	(1.468)		(4.611)	(1.565)
VIX (t-1)		0.202**	0.235***		0.211**	0.239***
		(2.412)	(3.054)		(2.427)	(2.897)
Industrial Production (t-1)		-0.377***	-0.337**		-0.392***	-0.350**
		(-4.699)	(-2.378)		(-5.040)	(-2.448)
CAY (q-1)		-0.519***	-0.579***		-0.529***	-0.591***
		(-4.229)	(-6.693)		(-4.345)	(-6.982)
Short Term Interest (t-1)			0.306			0.231
			(0.227)			(0.171)
Term Spread (t-1)			-0.735***			-0.709***
			(-3.887)			(-3.887)
Default Spread (t-1)			0.368**			0.412***
			(2.270)			(2.595)
Constant	0	0.04	0.506	0	0.043	0.352
	(-0.000)	(0.471)	(0.793)	(-0.000)	(0.515)	(0.526)
Observations	72	72	72	72	72	72
Adjusted R^2	0.362	0.739	0.781	0.338	0.737	0.777

Table 18 reports the results of the determinant of aggregate social-media disagreement. Column (1) to (3) reports the results using *Agg. Disagree* as an independent variable. Column (1) shows that 1 standard deviation increase in earning-price ratio *EP* predicts 0.520 future disagreements negatively and is significant at 5% level. Intuitively if current period earnings are higher, people tend to agree in the next period. However, the dividend-price ratio *DP* is founded to be positively predicting disagreement, which is counter-intuitive. Column (2) adds additional controls for some macroeconomic variables. For instance, a higher VIX represents the market's expectation of volatility, known as the fear index. In line with this notion, one standard deviation increase in VIX increases future aggregate social-media disagreement by 0.202 and is significant at a 5% level. Higher *Industrial Production* and *Inflation* represent a more optimistic growth in the economy and hence lower disagreement. My regression result also confirms that it is highly significant at the 1% level.

Moreover, Lettau and Ludvigson (2001) found that the Consumption-Wealth ratio *CAY* is a strong predictor of U.S. stock returns and is a proxy for a strong economy. My result shows a one standard deviation increase in *CAY* reduces future social media disagreement by 0.519 and is significant at 1% level, in line with the expectation that a consumption signifies a stronger economy leading to an increase in agreement in the stock market in the future. Lastly, column (3) adds the interest rate-related terms into the regression. Notably, the construction of *Term Spread* implies that the larger the spread, the 10-year long maturity treasury bond yield will have higher return for the same level of risk than short-term bond yield. This suggests a strong economy, consistent with my expectation that it predicts lower future aggregate social-media investor disagreement. Column (4) to (6) presents the result on the compressed *Agg. Disagree* using $0.5\,\beta + 0.5$ as the weight for stock level social-media disagreement. My result is robust and identical using such measures. Overall, my finding indicates that prior studies of macro-economic variables can predict and are related to aggregate social-media disagreement.

5.5 Aggregate disagreement and the SML line

To thoroughly examine how the aggregate social-media disagreement affects the slope of the securities market line, this section presents the relationships between the return and the SML line beta. Further, it categorizes this relationship during the months with the lowest (highest) aggregate social-media disagreement. As shown in Table 16, the stock-level social media disagreement is almost linearly increasing as the portfolio's beta increases. It is interesting to look at how this varies with low (high) aggregate social-media disagreements month.

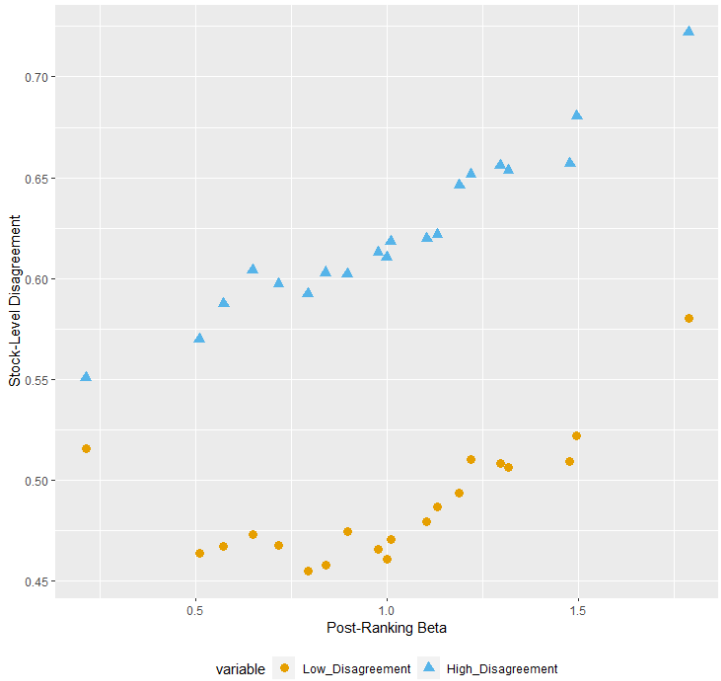
Furthermore, the model prediction of Hong and Sraer (2016)'s prediction is tested whereby (p. 2106) "*In low-disagreement months, the Security Market Line is upward-sloping. In high-disagreement months, the Security Market Line has a kink. Its slope is strictly positive for low-beta assets but strictly lower (and potentially negative) for high-beta assets. The Security Market Line should be more concave following months with high aggregate disagreement; equivalently, a portfolio long low- and high-beta assets and short medium-beta assets should experience lower performance following months of high aggregate disagreement.*"

5.5.1 Graphical analysis

Figure 12 below presents the relationships between stock-level social-media disagreement and beta-sorted portfolio during high aggregate social-media disagreement months (blue triangles) and low months (orange circle), defined as the top and bottom quartile of aggregate social-media disagreement. During high and low aggregate social-media disagreement months, both stock-level social-media disagreements increase as portfolio beta increases. In particular, investors disagree more in high aggregate social-disagreement months, supporting the fact that high beta stocks are riskier, and investors disagree more. One shall also note that aggregate social-media disagreement has no effects on the slope of both series. Indeed, there is no prior reason that during high aggregate social-media disagreement month, higher beta stocks will disagree (bending up) more than high beta stocks in low months. This is because when the overall market level disagreement increases, every stock experience the same during the month.

Figure 12
Stock-level social-media disagreement and beta-sorted portfolio

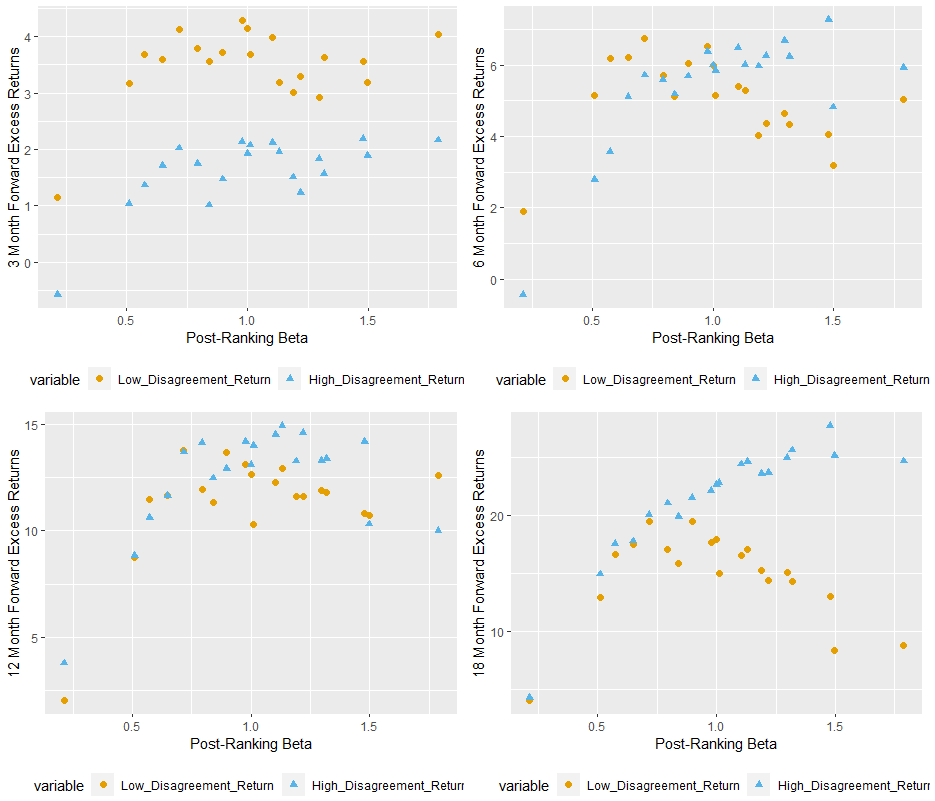
This figure shows the relationships between stock-level social-media disagreement and post-ranking beta from 12/2010 to 12/2017. In each calendar month stocks are ranked in increasing order with their estimated beta at the end of previous month, this is estimated using market model with 5 lags of market return. The ranked stocks are then sorted into 20 portfolios based on NYSE breakpoints. Post-ranking beta is obtained by regressing full sample of portfolio return on contemporaneous and 5 lags of market return. The graph plots the value-weighted average stock-level social-media disagreement in the 20 beta-sorted portfolio for months in the bottom quartile of aggregate social-media disagreement (orange circle) and months in the top quartile of aggregate social-media disagreement (blue triangle).



In addition, to understand how stock-level disagreement varies as portfolio-beta increases, figure 13 plots the relationships between average excess forward return and portfolio beta during low (orange circle) and high disagreement month (blue triangle), defined as the bottom and top quartile of aggregate social-media disagreement. For the return horizon, I followed Hong and Sraer (2016) using the average value-weighted excess forward return of 3 months (Top-Left), 6 months (Top-Right), 12 months (Bottom-Left), and 18 months (Bottom-Right).

Figure 13
Relationships Between Forward Excess Return, Beta, and Aggregate Social-Media Disagreement

This figure shows the relationships between stock-level social-media disagreement and post-ranking beta from 12/2010 to 12/2017. In each calendar month stocks are ranked in increasing order with their estimated beta at the end of previous month, this is estimated using market model with 5 lags of market return. The ranked stocks are then sorted into 20 portfolios based on NYSE breakpoints. Post-ranking beta is obtained by regressing full sample of portfolio return on contemporaneous and 5 lags of market return. The graph plots the average forward excess return over the next 3 months (Top-Left), 6 months (Top-Right), 12 months (Bottom-Left), and 18 months (Bottom-Right) of the 20 beta-sorted portfolios for months in the bottom quartile of aggregate disagreement (orange circle) and months in top quartile of aggregate disagreement (blue triangle). Aggregate disagreement is the monthly beta-weighted average of stock-level social-media disagreement.

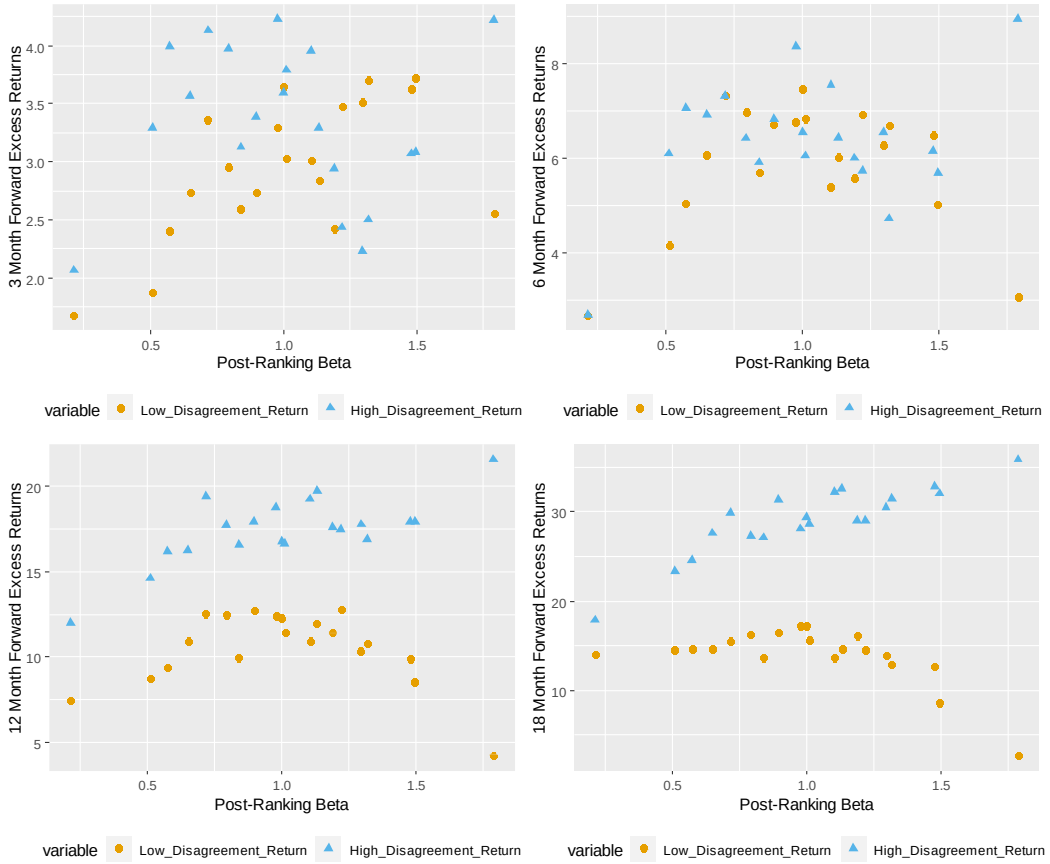


The 3 months' (Top-Left) relationships between forward excess return against the Post-Ranking Beta holds according to the theory that low aggregate social media disagreement months generate higher future return than in high aggregate social media disagreement months. Low disagreement months appear to be in a bullish or growth market; consistent with the literature, both SML lines are relatively flat. When the forward excess return is measured during the future 6, 12, 18 months, the SML line starts to curve, and the curvature is most noticeable at 12 months. One can also see that graph gets less noisy at 18 months, where the spread of each low and high disagreement months SML line is smaller. Interestingly, during high aggregate social media disagreement months, the return and beta relationships remain almost linearly increasing, whereas low disagreement months exhibit an inverted U-shape. The flipped relationships suggest Hong and Sraer's theory only holds for a short investment horizon in the sample period.
For robustness, I followed Hong and Sraer (2016) construction of Aggregate disagreement using the standard deviation of I/B/E/S analyst forecast of long-run EPS growth (EPS LTG).

Figure 14 below plots such relationships with the same sample period from 12/2010 to 12/2017 as in Figure 13. The low and high disagreement pattern of the forward excess returns at 3 (Top-Left) and 6 months (Top-Right) are less visible than our aggregate social-media disagreement. Secondly, it is rather surprisingly that the forward excess returns at 12 (Bottom-Left) and 18 months (Bottom-Right) during the high disagreement months (blue triangle) have a higher return increasing linearly with higher beta portfolio. Note that the low disagreement months (orange circle) exhibit a nonlinear inverted U shape pattern, this is consistent with aggregate disagreement generated through social media. The reason why Hong and Sraer (2016)'s theory does not hold is believed to be attributed to the chosen sample period and market structure changed. Their original paper was based on the period from 12/1981 to 12/2014, containing 381 months, while the sample in question only has 84 months. In unreported results, I confirmed and successfully replicated their result using the original sample period.

Figure 14
The Relationships Between Forward Excess Return, Beta, and Aggregate Disagreement using Hong and Sraer (2016) Original Construction

This figure shows the relationships between stock-level disagreement and post-ranking beta from 12/2010 to 12/2017. In each calendar month, stocks are ranked in increasing order with their estimated beta at the end of the previous month. This is estimated using market model with 5 lags of market return. The ranked stocks are then sorted into 20 portfolios based on NYSE breakpoints. Post-ranking beta is obtained by regressing the full sample of portfolio return on contemporaneous and 5 lags of market return. The graph plots the average forward excess return over the next 3 months (Top-Left), 6 months (Top-Right), 12 months (Bottom-Left), and 18 months (Bottom-Right) of the 20 beta-sorted portfolios for months in the bottom quartile of aggregate disagreement (orange circle) and months in top quartile of aggregate disagreement (bule triangle). Aggregate disagreement is the monthly beta-weighted average of stock-level disagreement measured as the standard deviation of analyst forecast of long-run EPS growth.



5.5.2 Regression analysis

Figure 13 reveals that the concavity of the SML line is much clearer in forward 12- and 18-months forward return, and such effect only appears in low aggregate social-media disagreement months. Since the graphical analysis provides a limited conclusion, a formal test of Hong and Sraer (2016) prediction defined in section 5.5 is conducted using a two-stage Fama and MacBeth (1973) style regression and also according to De Giorgi et al. (2019). For each month, the cross-sectional regression is estimated over the 20 beta-sorted portfolios:

$$r^{(k)}_{P,t} = \kappa_t + \pi_t \beta_P + \varphi_t \beta_P^2 + \varepsilon_{P,t}, \text{ where } P = 1, \ldots, 20 \text{ and } k = 6, 9, 12, 18 \qquad (12)$$

where $r^{(k)}_{P,t}$ is the k-month forward excess return of P^{th} beta-sorted portfolio and β_P is the full-sample post ranking beta of P^{th} beta-sorted portfolio.

The coefficients κ_t, π_t, and φ_t are obtained by following Hong and Sraer (2016) in their choices of k-month and using the specification in equation 12. Although the model is a single period model, the choice of k should be limited to one day or month. One may argue that the investment horizon also matters as trading friction prevents stocks from being traded frequently; hence the result shows different choices of k for comparison.

By running the cross-section regression, it fits a concave line on 20 betas sorted portfolios each month, and the coefficients represent the SML line. Note that this method might suffer from error-in-dependent-variable where the coefficients of π_t and φ_t may not have control over the other confounding factors, which is believed to be one of the downsides of the analysis.

Once a time-series of all the coefficients are obtained, they are regressed on several macro-economic factors that could explain the concavity of the Securities Market Line. For instance, *Ted Spread* can affect the slope of the SML line, according to Frazzini and Pederson (2014). The coefficient term φ_t in the β_P^2 term will be of primary interest here as it captures the concavity of the Securities Market Line. φ_t can also be considered the profit using a strategy of going long the bottom and top beta portfolios of ({1,2}, {19,20}) and short all other portfolios (3,…,18).

The list of macro-economic factors includes *Agg. Disp.* is the beta-weighted average stock-level social-media disagreement as calculated in section 5.3.2 at month *t-1*, including also the standard Fama and French (1992) risk factors, namely $R_m^{(k)}(t)$, $HML^{(k)}(t)$, $SMB^{(k)}(t)$, and Jegadeesh and Titman (1993) momentum factor $UMD^{(k)}(t)$. Where k represents the return period of $k=6, 9, 12, 18$, and the return factors are cumulated according to the period in question. These

standard risk factors have an incremental explanation to Securities Market Line. Cohen, Polk, and Vuolteenaho (2005) found that the inflation rate *Inflation (t-1)* measured by the year-over-year inflation in month *t-1* and market-wide dividend-to-price ratio *D/P (t-1)* at month *t-1* has explanatory power in the cross-section of return, I nonetheless controlled for them in regression. Moreover, Frazzini and Pedersen (2014) suggested that *Ted Spread (t-1)* is a proxy for an investor's funding constraint or credit risk. When Ted Spread increases, it reflects an increase in interbank lending rate leading to a tightening investor funding constraint. Investors are less likely to borrow money from banks to short-sell overpriced stocks, which constitutes a flatter or downward sloping Securities Market Line. *VIX (t-1)* is the CBOE volatility index and in part explains the disagreement in the stock market according to Barinov (2013), *CAY (t-1)* is the consumption-wealth ratio obtained from Martin Lettau's website, which captures time-varying risk premium. *Short Term Interest (t-1)* is the 3-month money market rate where Campbell (1987) shows it can predict future stock returns, *Default Spread (t-1)* is the spread between Baa corporate bond minus 10-year maturity bond, and lastly, *Term Spread (t-1)* is the 10-year treasury minus 3-month treasury. All variables are standardized and standard errors are adjusted for serial correlation using Newey West (1987) with 12 lags.

TABLE 19
Regression Analysis on Concavity of the Securities Market Line (Value-Weighted)

This table shows the regression analysis on concavity of the Securities Market Line. The sample uses CRSP stock return tape and exclude penny stocks (price < $5) and microcaps stocks defined as the bottom two deciles using NYSE stocks as break points. Every calendar month stocks are pre-ranked according to their estimated beta using daily return over past calendar year on 5 lags of market model. The ranked stocks are sorted into 20 buckets to form 20 beta-sorted portfolios using NYSE stocks as break points. The post-ranking portfolio beta is then computed using full-sample and with the same market model.

The regression analysis employs a two-stage Fama and Macbeth (1973) style regression where I estimate the first stage cross sectional regression:

$$r^{(k)}_{P,t} = \kappa_t + \pi_t \beta_P + \varphi_t \beta_P^2 + \varepsilon_{P,t}, \text{ where } P = 1, \ldots, 20 \text{ and } k = 6, 9, 12, 18$$

where $r^{(k)}_{P,t}$ is the k-month value-weighted forward excess return of the P^{th} beta-sorted portfolio and β_P is the full-sample post-ranking beta of the P^{th} beta-sorted portfolios. I retrieved the coefficient estimates of κ_t, π_t, φ_t each month and regress on a time-series regression in second stage:

Panel A: $\varphi_t = \alpha + \mu_1 Agg.\ Disp_{t-1} + \mu_2 R_{m,t}^{(k)} + \mu_3 HML_t^{(k)} + \mu_4 SMB_t^{(k)} + \mu_5 UMD_t^{(k)} + \mu_x \boldsymbol{X}_{t-1} + \varepsilon_t$
Panel B: $\pi_t = \alpha + \mu_1 Agg.\ Disp_{t-1} + \mu_2 R_{m,t}^{(k)} + \mu_3 HML_t^{(k)} + \mu_4 SMB_t^{(k)} + \mu_5 UMD_t^{(k)} + \mu_x \boldsymbol{X}_{t-1} + \varepsilon_t$

Panel A, B presents the results for φ_t, π_t, κ_t respectively and various forward *k*-month excess return for comparison. Column (1), (3), (5), (7) presents the results with Hong and Sraer (2016)'s list of controls where as column (2), (4), (6), (8) include additional controls.

Agg. Disp. is the beta-weighted average stock-level social-media disagreement as calculated in section 5.3.1. $R_m^{(k)}$, $HML^{(k)}$, $SMB^{(k)}$ are the cumulative k-months market, high-minus-low, small-minus-big return factor respectively. $UMD^{(k}$ is the cumulative k-months up-minus-down momentum factor. *k* represents the return period of *k=6, 9, 12, 18*. *Inflation* is the year-over-year inflation. *D/P* is the Dividend-Price ratio obtained from Robert Shiller's website. *TED Spread* is the difference between three-month LIBOR based in US dollars and three-month Treasury bill. *VIX* is the CBOE volatility index. *CAY* is consumption wealth ratio obtained from Martin Lettau website. *Short-term interest* obtained from OECD website. *Default Spread* is the Baa corporate bond minus 10-year maturity treasury bond and lastly *Term Spread* is the 10-year treasury minus 3-month treasury. Standard errors are adjusted using Newey West (1987) to correct for serial-correlation using 12 lags. t-statistics are in parenthesis and ***, **, * denotes 1%, 5%, 10% significance respectively.

Panel A: $\boldsymbol{\varphi_t}$

	Forward k=6-month excess return		Forward k=9-month excess return		**Forward k=12-month excess return**		Forward k=18-month excess return	
Dep. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Agg. Disp. (t-1)	-0.392	-3.893**	-4.730***	-3.196	-6.583***	-4.322*	-7.812***	-6.433**
	(-0.198)	(-2.171)	(-3.465)	(-1.471)	(-3.971)	(-1.736)	(-3.204)	(-2.180)
Rm(k) (t)	-0.246	-0.324**	-0.338*	-0.264	-0.309***	-0.310***	-0.398***	-0.288*
	(-1.514)	(-1.979)	(-1.856)	(-1.401)	(-3.244)	(-2.832)	(-2.916)	(-1.830)
HML(k) (t)	0.309	0.143	0.588***	0.533***	0.514***	-0.305*	0.518***	0.025
	(1.124)	(0.698)	(3.143)	(3.084)	(2.991)	(-1.716)	(2.667)	(0.096)
SMB(k) (t)	0.854***	0.541**	1.055***	1.072***	0.904***	0.840***	1.155***	0.882***
	(4.041)	(2.543)	(4.828)	(5.705)	(3.133)	(3.125)	(2.654)	(3.822)
UMD(k) (t)	-0.141	-0.206	-0.11	-0.045	-0.032	-0.520*	-0.319**	-0.475**
	(-0.701)	(-0.794)	(-0.999)	(-0.170)	(-0.296)	(-1.885)	(-2.262)	(-2.049)
DP (t-1)	-0.002	-1.034	-1.132	0.019	1.813	0.656	3.889**	1.645
	(-0.001)	(-0.338)	(-0.521)	(0.007)	(0.746)	(0.283)	(2.547)	(1.116)
Inflation (t-1)	0.278	2.578**	-0.754	0.473	-0.553	3.101**	0.03	4.540***
	(0.177)	(2.001)	(-0.524)	(0.467)	(-0.549)	(2.369)	(0.038)	(3.648)
Ted Spread (t-1)	-1.281	-2.925	1.294	3.903***	0.386	6.041***	-1.697	2.944
	(-0.596)	(-1.196)	(0.828)	(3.568)	(0.209)	(3.705)	(-1.528)	(1.438)
Vix (t-1)		-1.670***		-2.501**		-2.195***		-2.510***
		(-2.713)		(-2.146)		(-2.849)		(-2.670)
Cay (last quarter)		-2.595**		-1.096		-1.489		-4.918***
		(-2.200)		(-0.611)		(-0.632)		(-2.854)
Short Term Interest (t-1)		1.884		-3.710**		-10.564***		-6.967***
		(0.865)		(-2.139)		(-5.312)		(-3.481)
Term Spread (t-1)		-3.952		1.912		-3.683		-2.236
		(-1.400)		(0.517)		(-1.238)		(-1.255)
Default Spread (t-1)		2.662		1.287		4.559**		5.366***
		(1.449)		(0.599)		(2.160)		(3.120)
Constant	-2.422	-6.251***	-2.543	-5.150*	-5.728***	-9.602***	-7.506***	-15.930***
	(-0.836)	(-3.315)	(-1.160)	(-1.923)	(-3.247)	(-2.999)	(-5.596)	(-4.988)
Observations	78	78	75	75	72	72	66	66
Adjusted R^2	0.223	0.341	0.445	0.469	0.451	0.595	0.422	0.535

TABLE 19 (Cont.)

Panel B: π_t

	Forward k=6-month excess return		Forward k=9-month excess return		**Forward k=12-month excess return**		Forward k=18-month excess return	
Dep. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Agg. Disp. (t-1)	1.974	5.378*	8.775***	3.012	9.753***	1.707	17.360***	9.510*
	(0.529)	(1.903)	(2.958)	(0.713)	(2.979)	(0.467)	(3.914)	(1.847)
$Rm^{(k)}$ (t)	1.489***	1.431***	1.653***	1.175***	1.258***	0.885***	1.403***	1.401***
	(4.010)	(3.639)	(5.384)	(2.978)	(6.601)	(2.838)	(6.025)	(6.387)
$HML^{(k)}$ (t)	0.102	0.508	-0.433	-0.032	-0.849***	1.041**	-0.877***	-0.072
	(0.173)	(1.290)	(-1.030)	(-0.097)	(-2.903)	(2.220)	(-2.614)	(-0.209)
$SMB^{(k)}$ (t)	-0.567	0.114	-0.815**	-0.942***	-0.584	-0.759	-0.668	-0.282
	(-1.237)	(0.218)	(-2.008)	(-2.835)	(-1.112)	(-1.588)	(-0.927)	(-0.622)
$UMD^{(k)}$ (t)	0.575	0.558	0.641**	0.302	0.344	0.897	1.002***	1.061***
	(1.538)	(1.297)	(2.339)	(0.549)	(1.371)	(1.600)	(4.759)	(2.874)
DP (t-1)	-2.251	-0.609	-1.384	-2.239	-1.06	0.611	-2.885	-1.4
	(-0.535)	(-0.128)	(-0.311)	(-0.422)	(-0.276)	(0.171)	(-1.486)	(-0.644)
Inflation (t-1)	-2.863	-5.942***	-1.743	-0.655	-0.439	-3.262	1.044	-3.969
	(-0.911)	(-3.009)	(-0.646)	(-0.382)	(-0.234)	(-0.877)	(0.667)	(-1.625)
Ted Spread (t-1)	1.236	5.207	-2.115	-8.027***	0.603	-15.291***	2.172	-3.329
	(0.320)	(1.430)	(-0.663)	(-3.416)	(0.191)	(-7.875)	(0.995)	(-1.076)
Vix (t-1)		6.043***		5.391**		4.337***		4.738**
		(4.825)		(2.419)		(2.577)		(2.264)
Cay (last quarter)		2.129		-0.267		2.351		2.533
		(1.102)		(-0.101)		(0.700)		(0.757)
Short Term Interest (t-1)		-4.334		7.033**		24.453***		13.023***
		(-1.149)		(2.100)		(6.292)		(3.991)
Term Spread (t-1)		4.134		-8.06		-0.998		0.378
		(0.763)		(-0.952)		(-0.151)		(0.149)
Default Spread (t-1)		-7.897**		-6.893		-13.090***		-12.592***
		(-2.378)		(-1.532)		(-3.408)		(-3.186)
Constant	4.185	10.516***	2.598	5.176	7.972	13.471***	6.746***	11.321**
	(0.648)	(2.804)	(0.485)	(1.038)	(1.467)	(2.676)	(3.146)	(2.334)
Observations	78	78	75	75	72	72	66	66
Adjusted R^2	0.265	0.43	0.359	0.417	0.323	0.591	0.549	0.63

Table 19 panel A shows the time-series regression result by regressing the squared term coefficient φ_t from the cross-sectional regression on a list of macro-economic variables using the value-weighted return portfolio. Odd number columns include a list of explanatory variables used in Hong and Sraer (2016) analysis, whereas even number columns include explanatory variables that are hypothesized to have explanatory power. Across all odd columns, the explanatory variable of interest *Agg. Disp. (t-1)* is negative and significant at 1% level other than forward k = 6-month excess return and is consistent with our graphical analysis in Figure 13 that there is a noisy pattern at $k = 6$. The significance of *Agg. Disp. (t-1)* confirms my finding and Hong and Sraer (2016)'s theory that the aggregate disagreement constructed using social media explains the increasing concavity of the Securities Market Line.

Focusing in column (5) where forward $k = 12$-month return is the standard time horizon in literature, small-minus-big factor $SMB^{(12)}$ *(t)* explains less concave SML line at 1% significance level. Inconsistent with prior literature that higher return from small firms are riskier than large firms, where they have lower current price and higher expected return, higher disagreement and thus more concave the SML line.

Interestingly *Ted Spread (t-1)* does not significantly explain the concavity of the SML line as opposed to Frazzini and Petersen (2014) findings. Column (6) includes more controls; the social media aggregate disagreement still predicts a more concave SML line albeit at a lower significance of 10% level. *Ted Spread (t-1)* at this specification flipped sign surprisingly and significantly predicts a less concave SML line at 1% level. *VIX (t-1),* known as the "fear gauge", measures the market's expectation on S&P 500 index option price and can be viewed as overall market uncertainty. Intuitively higher uncertainty spans higher disagreement and hence a more concave SML line.

Short-term interest (t-1) finds positive and significant relationships that higher short-term interest predicts a more concave SML line. A higher short-term interest rate reduces the investor's ability to borrow money and short-sell stocks. More investor is sidelined from shorting high disagreement over-priced stocks, leading to a lower future return. Panel B regresses π_t on various macroeconomic variables, and it shows the rate of change of the curvature of the Securities Market Line. In all settings, social-media disagreement predicts an increasing convex rate of SML line while only the return horizon in forward 9-, 12-, 18-month excess return is significant using Hong and Sraer (2016)'s series of the explanatory variable.

TABLE 20
Regression Analysis on Concavity of the Securities Market Line (Equal-Weighted)

This table shows the regression analysis on concavity of the Securities Market Line. The sample uses CRSP stock return tape and exclude penny stocks (price < $5) and microcaps stocks defined as the bottom two deciles using NYSE stocks as break points. Every calendar month stocks are pre-ranked according to their estimated beta using daily return over past calendar year on 5 lags of market model. The ranked stocks are sorted into 20 buckets to form 20 beta-sorted portfolios using NYSE stocks as break points. The post-ranking portfolio beta is then computed using full-sample and with the same market model.

The regression analysis employs a two-stage Fama and Macbeth (1973) style regression where I estimate the first stage cross sectional regression:

$$r^{(k)}_{P,t} = \kappa_t + \pi_t \beta_P + \varphi_t \beta_P^2 + \varepsilon_{P,t}, \text{ where } P = 1, \ldots, 20 \text{ and } k = 6, 9, 12, 18$$

where $r^{(k)}_{P,t}$ is the k-month equal-weighted forward excess return of the P^{th} beta-sorted portfolio and β_P is the full-sample post-ranking beta of the P^{th} beta-sorted portfolios. I retrieved the coefficient estimates of κ_t, π_t, φ_t each month and regress on a time-series regression in second stage:

Panel A: $\varphi_t = \alpha + \mu_1 Agg.\ Disp_{t-1} + \mu_2 R_{m,t}^{(k)} + \mu_3 HML_t^{(k)} + \mu_4 SMB_t^{(k)} + \mu_5 UMD_t^{(k)} + \mu_x \boldsymbol{X}_{t-1} + \varepsilon_t$
Panel B: $\pi_t = \alpha + \mu_1 Agg.\ Disp_{t-1} + \mu_2 R_{m,t}^{(k)} + \mu_3 HML_t^{(k)} + \mu_4 SMB_t^{(k)} + \mu_5 UMD_t^{(k)} + \mu_x \boldsymbol{X}_{t-1} + \varepsilon_t$
Panel C: $\kappa_t = \alpha + \mu_1 Agg.\ Disp_{t-1} + \mu_2 R_{m,t}^{(k)} + \mu_3 HML_t^{(k)} + \mu_4 SMB_t^{(k)} + \mu_5 UMD_t^{(k)} + \mu_x \boldsymbol{X}_{t-1} + \varepsilon_t$

Panel A, B, C presents the results for φ_t, π_t, κ_t respectively and various forward *k*-month excess return for comparison. Column (1), (3), (5), (7) presents the results with Hong and Sraer (2016)'s list of controls where as column (2), (4), (6), (8) include additional controls.

Agg. Disp. is the beta-weighted average stock-level social-media disagreement as calculated in section 5.3.1. $R_m^{(k)}$, $HML^{(k)}$, $SMB^{(k)}$ are the cumulative k-months market, high-minus-low, small-minus-big return factor respectively. $UMD^{(k}$ is the cumulative k-months up-minus-down momentum factor. *k* represents the return period of *k=6, 9, 12, 18. Inflation* is the year-over-year inflation. *D/P* is the Dividend-Price ratio obtained from Robert Shiller's website. *TED Spread* is the difference between three-month LIBOR based in US dollars and three-month Treasury bill. *VIX* is the CBOE volatility index. *CAY* is consumption wealth ratio obtained from Martin Lettau website. *Short-term interest* obtained from OECD website. *Default Spread* is the Baa corporate bond minus 10-year maturity treasury bond and lastly *Term Spread* is the 10-year treasury minus 3-month treasury. Standard errors are adjusted using Newey West (1987) to correct for serial-correlation using 12 lags. t-statistics are in parenthesis and ***, **, * denotes 1%, 5%, 10% significance respectively.

Panel A: $\boldsymbol{\varphi_t}$

	Forward k=6-month excess return		Forward k=9-month excess return		**Forward k=12-month excess return**		Forward k=18-month excess return	
Dep. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Agg. Disp. (t-1)	0.317	-0.903	-1.323	0.071	-0.27	0.312	0.263	-2.975
	(0.289)	(-1.140)	(-1.561)	(0.082)	(-0.341)	(0.420)	(0.330)	(-1.582)
$Rm^{(k)}$ (t)	0.057	-0.078	-0.083	-0.177	-0.03	-0.104	-0.09	-0.073
	(0.539)	(-0.842)	(-0.705)	(-1.453)	(-0.284)	(-1.243)	(-1.257)	(-0.870)
$HML^{(k)}$ (t)	-0.004	0.092	-0.161	0.077	-0.246***	-0.126	0.004	0.145
	(-0.036)	(0.766)	(-1.574)	(0.891)	(-2.905)	(-0.877)	(0.021)	(0.698)
$SMB^{(k)}$ (t)	0.689***	0.707***	0.655***	0.694***	0.310**	0.282**	0.151	-0.024
	(4.890)	(6.762)	(4.556)	(6.350)	(2.309)	(2.310)	(0.773)	(-0.127)
$UMD^{(k)}$ (t)	0.071	0.103	0.101*	0.140**	0.166**	0.224**	-0.05	-0.105
	(0.769)	(1.080)	(1.748)	(2.139)	(2.082)	(2.548)	(-0.533)	(-0.748)
DP (t-1)	0.015	1.799*	1.710*	4.179***	4.377***	5.555***	5.802***	5.130***
	(0.014)	(1.885)	(1.917)	(3.552)	(5.683)	(3.916)	(4.894)	(3.795)
Inflation (t-1)	0.321	0.257	0.065	0.319	0.478	0.433	-0.008	1.664*
	(0.495)	(0.381)	(0.128)	(0.423)	(0.939)	(0.401)	(-0.011)	(1.779)
Ted Spread (t-1)	-2.286***	-3.397***	-1.420***	-1.477	-3.079***	-4.454***	-5.566***	-5.337***
	(-3.020)	(-3.204)	(-3.994)	(-1.297)	(-5.948)	(-3.228)	(-4.544)	(-6.184)
Vix (t-1)		-0.131		-1.366***		-1.079		-1.452**
		(-0.391)		(-2.726)		(-1.416)		(-2.094)
Cay (last quarter)		0.325		1.166		1.161		-2.864**
		(0.430)		(1.147)		(0.815)		(-2.298)
Short Term Interest (t-1)		2.542***		-0.122		1.67		1.847***
		(3.138)		(-0.104)		(1.026)		(2.965)
Term Spread (t-1)		0.774		2.197*		0.872		-1.761
		(0.526)		(1.803)		(0.467)		(-1.232)
Default Spread (t-1)		-1.073		-2.112*		-0.516		-0.622
		(-1.370)		(-1.847)		(-0.424)		(-0.917)
Constant	-5.945***	-5.131***	-7.384***	-6.655***	-10.800***	-9.757***	-12.467***	-16.980***
	(-5.743)	(-5.268)	(-7.474)	(-4.943)	(-13.514)	(-5.066)	(-10.822)	(-8.155)
Observations	78	78	75	75	72	72	66	66
Adjusted R^2	0.462	0.539	0.443	0.515	0.432	0.419	0.51	0.614

TABLE 20 (Cont.)

Panel B: π_t

	Forward k=6-month excess return		Forward k=9-month excess return		**Forward k=12-month excess return**		Forward k=18-month excess return	
Dep. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Agg. Disp. (t-1)	0.526	0.583	2.662	-2.184	-0.916	-5.218**	1.375	3.584
	(0.247)	(0.395)	(1.601)	(-1.186)	(-0.541)	(-2.514)	(0.702)	(0.932)
$Rm^{(k)}$ (t)	0.770***	0.912***	0.984***	1.000***	0.687***	0.654***	0.727***	0.831***
	(3.258)	(3.609)	(3.315)	(2.699)	(2.877)	(3.172)	(3.453)	(2.974)
$HML^{(k)}$ (t)	0.428**	0.256	0.739***	0.431**	0.673***	0.563*	0.268	-0.273
	(2.168)	(1.211)	(3.204)	(2.347)	(3.017)	(1.651)	(1.051)	(-0.717)
$SMB^{(k)}$ (t)	-0.476*	-0.496***	-0.331	-0.430**	0.419	0.255	1.380***	1.483***
	(-1.911)	(-3.123)	(-1.495)	(-2.563)	(1.423)	(0.790)	(2.699)	(3.312)
$UMD^{(k)}$ (t)	0.014	-0.156	0.036	-0.183	-0.074	-0.401	0.461***	0.309
	(0.066)	(-0.774)	(0.253)	(-1.302)	(-0.494)	(-1.082)	(2.779)	(1.158)
DP (t-1)	-1.286	-5.455***	-5.616***	-10.266***	-7.697***	-11.269***	-8.833***	-11.383***
	(-0.808)	(-4.119)	(-3.080)	(-3.825)	(-3.509)	(-2.684)	(-3.406)	(-2.824)
Inflation (t-1)	-1.602	-0.406	-1.596	-0.006	-1.613	1.01	2.378	2.551
	(-1.134)	(-0.465)	(-1.249)	(-0.004)	(-1.180)	(0.462)	(1.293)	(1.296)
Ted Spread (t-1)	3.114**	5.748***	2.800***	2.987	6.802***	6.699**	8.992***	11.434***
	(2.179)	(3.901)	(3.201)	(1.196)	(4.586)	(2.169)	(4.479)	(6.261)
Vix (t-1)		1.966***		3.557***		2.816*		3.567
		(2.634)		(3.098)		(1.729)		(1.634)
Cay (last quarter)		-2.744		-4.366**		-3.719		-0.166
		(-1.464)		(-2.464)		(-1.519)		(-0.049)
Short Term Interest (t-1)		-5.565***		-0.374		-0.849		-4.552**
		(-4.504)		(-0.168)		(-0.215)		(-2.410)
Term Spread (t-1)		-4.368*		-7.342***		-7.27		-1.882
		(-1.673)		(-2.867)		(-1.377)		(-0.578)
Default Spread (t-1)		0.798		1.567		-0.551		0.621
		(0.592)		(0.713)		(-0.227)		(0.248)
Constant	10.097***	7.394***	11.604***	8.339***	17.310***	13.170***	15.033***	14.671**
Agg. Disp. (t-1)	(3.464)	(3.278)	(4.676)	(3.333)	(7.048)	(3.446)	(5.177)	(2.365)
Observations	78	78	75	75	72	72	66	66
Adjusted R^2	0.448	0.569	0.508	0.587	0.509	0.514	0.727	0.736

Table 20 shows the regression result with an equally weighted portfolio. The Panel A of table 20 consists of the same test as that of table 19. Under this setting, my aggregate social-media disagreement was not significant at all return horizon. Some of the coefficients were even fluctuating between positive and negative in different return periods. However, *Ted Spread (t-1)* is consistent with the theory that it significantly predicts a flatter or convex SML line. For instance, at *k*=12 month, it predicts -4.454% (6.699%) of φ_t (π_t) respectively. A possible explanation could be that the value-weighted portfolio firms which are large in market capitalization experience the highest portion of aggregate social-media disagreement. An investor who wishes to short-sell these stocks is sidelined, especially in high-beta, large market capitalization stocks. It channels through the fact that these stocks are over-priced. They are inherently expensive and are above what individual investors could afford, together with aggregate social-media disagreement from the market generating a lower future return. Whereas every stock's portfolio return is averaged out equally, aggregate social-media disagreement does not matter much to the high market capitalization firms.

For *D/P*, it is consistently and positively explaining φ_t which is the curvature of SML line. For instance, at *k=9* forward return interval, D/P explains 4.179% (-10.266%) of φ_t (π_t) and both are significant at 1% level. This positive relationship in curvature at decreasing rate suggests that higher dividend price ratio could make SML line more convex like. The likely explanation would be that a company with higher dividend price ratio could provide more money to return to its shareholders. If more dividends are paying out, investors would likely hold the stock rather than trading and short-selling the security. In effect, it reverts the concavity of the SML line. Compared with the value-weighted portfolio in Table 19, the Dividend yield is also more significant, given an equal influence in return for stocks in an equally weighted portfolio.

5.6 Summary for Chapter 5

This chapter aims to study the long-standing problem that spans several decades in testing the CAPM model empirically, where higher risks do not deliver higher returns. Graphically speaking, when plotting expected return versus beta risk, the empirical Securities Market Line is downward sloping or even concave shape. Several academics propose that investors are limited by short-sell constraint whereby over-priced stocks, as appeared in those high beta stocks, are sidelined from short-selling, leading to a lower future return. A trading strategy called Bet Against Beta (BAB) was proposed by Frazzini and Pedersen (2014) to short-sell high-beta stocks until the portfolio reaches beta of 1, and use the proceeds to purchase low-beta stocks to create a zero-cost self-sufficient portfolio.

Short selling constraint partially explains the downward sloping Securities Market Line. However, it does not explain the concavity of the SML line in some recent empirical research, for example, argued by De Giorgi, Post, and Yalcin (2020). Precisely why did low beta risk stock appear to follow the CAPM model's prediction that it features a positive linear relationship of higher risk, higher return, and only high beta risk stock break such relationships? Hong and Sraer (2016) proposed including short-sell constraint and relaxing a fundamental assumption in the CAPM model that individual investors have homogeneous expectations. Their model believed that two groups of optimist and pessimist investors can disagree on the fundamental of a stock, although Low-beta stocks are less prone to disagreement due to their stable fundamentals such as a utility company.

In contrast, high-beta stocks are more subject to scrutiny and disagreement such as the start-up companies, which are more likely to have fundamental uncertainty. In the same vein, if a macroeconomic wide disagreement exists in the market, high-beta stocks are mostly affected since their prices are more volatile than the market, and their firm-specific risk is also the highest. Intuitively, when disagreement in both macro and firm-specific components are high, pessimistic investors are sidelined from short-selling high-beta stocks, leading to lower future returns.

This section contributes to the literature as among the first in using the social media, primarily non-institutional investors, as a proxy to investor disagreement. Retail investors are particularly short-sell constrained, and in theory, amplify the concavity of the SML line. More importantly, prior research proxies macro-disagreement from macro series like *Survey of*

Professional Forecaster. The social media disagreement in this study is created bottom-up from stock-level disagreement, rather than a measure of disagreement aggregated at firm-level.

The other contribution to the literature shows that this aggregate social media disagreement can be determined and predicted by various macroeconomic indicators, i.e. C*onsumption-wealth ratios*, *Dividend Price ratio*, *default spread*, and *term spread*, to name a few, consistent with the stock market activity. The social-media disagreement also reveals that stocks sorted into twenty portfolios using the beta, and averaging the stock-level social-media disagreement into portfolio-level, the portfolio-level social-media disagreement increases linearly with beta. The finding confirms that social-media investors disagree more on high-beta stocks.

The main finding in this section tested the hypothesis from Hong and Sraer (2016) that in low aggregate disagreement months, Securities Market Line remains a positive linear relationship of risk and expected return according to CAPM. In contrast, the concavity only appears in high aggregate disagreement months. Under graphical examination, I found that such hypothesis doesn't hold when high aggregate disagreement months experience positive linear relationships in forward future 9, 12, 18-month return. To ensure the Social-Media disagreement does not drive the inconsistent result, I replicated the Hong and Sraer (2016) original paper's disagreement using I/B/E/S analyst forecast dispersion and confirmed the findings. The inconsistency occurs due to the sample period from 2010 to 2017.

Lastly, a regression analysis was used to statistically detect whether aggregate social-media disagreement explains the concavity of the Securities Market Line and their relationships. The test follows Fama and Macbeth (1973) style regression in which forward future return is regressed on beta and beta squared in the cross-section and collected the coefficients. The regression is analogous to fitting a non-linear curve as the empirical Securities Market Line. The second stage regresses the time series coefficients on various macroeconomic factors that could explain the coefficients. In particular, the squared coefficient is of interest here because it captures the concavity of the Securities Market Line. Results suggest that aggregate social-media disagreement is robust in explaining the squared coefficient at forward future 9, 12, 18-month return. Although such a method is performed in various studies; however, it suffers from error-in-dependent-variable where the coefficients may not have controlled over other confounding factors, which could be one of the limitations of the analysis.

6. Conclusion

The finance literature has shifted from the traditional framework to incorporating human psychology in explaining financial anomalies. Traditional frameworks like the CAPM asset pricing model faced several criticisms and their empirical failure can mainly be attributed to the unrealistic assumptions that investors are rational and be able to process information immediately with ease. However, questions are not answered about how individual investors formulate a portfolio to implement their investment decisions, and why higher risk does not always deliver high returns. Behavioural finance, on the other hand, seeks to answer these questions from a more realistic body of knowledge using psychology based upon the followings: First, due to limited informational processing and other behavioural biases, investors often make systematically suboptimal and irrational decisions in trading; Second, real-world is not frictionless, which means limits to arbitrage exist and hinders rational arbitrageurs from correcting asset mispricing.

My research objectives explored individual investors' limited attention by leveraging the uniqueness of the tweets dataset at user-level ("investor" or "user" used interchangeably), and studied how investors' dispersion of beliefs with social media tweets disagreement affected the pricing in the cross-section of stocks return.

Recent literatures have been trying to address the deficiency of the CAPM assumptions, for example, by introducing behavioural finance on the asset pricing model. Yet so far the studies have been focused on discovering firm characteristics under firm-level analysis, which assumed individual investors acts homogeneously. What we don't know is the kind of relationships that exist on investors' decisions in the stock market, for instance the social-media sentiment and stock returns. To answer these questions, this thesis examines the behavioural bias faced by an individual investor, using a unique dataset of the StockTwits finance social-media platform to infer two main biases, namely limited attention and investor disagreement.

The approach is to extract information and classify the tweets from StockTwits into a quantitative dataset. To do so, literatures in computer science and artificial intelligence are studied to find a suitable tool for the job. "FastText", a current generation state-of-the-art machine learning algorithm, is chosen on its merits of being a black box over other quantitative software such as NVivo. Bojanowski et al. (2017), points out that this preferred method surpasses any other existing models regarding classification accuracy. "FastText" classifies the

tweets into {*Bullish, Neutral, Bearish*} category and created a *Sentiment* index whereas NVivo relies on a predefined dictionary and matches the words in StockTwits in each sentiment category (*Bullish, Bearish*} to arrive at a sentiment value. The latter approach suffers from undercounting due to many abbreviations and non-frequent word used in StockTwits, thereby underestimating the sentiment value.

Despite the limitation of using StockTwits as a single dataset, what it differentiates from other social-media platforms is that it also provides user-level information such as the number of stocks individual investors followed, number of likes received from fellow investors, etc., which allows running analysis at the user level. I am grateful to be amonst those who pioneer the research of behaviour finance at user-firm-time-level sentiment as the data granularity continue to grow. Given the rich information from this dataset, several policy implications can be drawn based on the findings presented in this thesis.

The task of investing in stocks requires significant time and effort in gathering, analysing, and interpreting financial information. While investors following multiple stocks are likely to have superior market-wide knowledge, those specializing in only a few have a comparative advantage in understanding in greater depth. Nevertheless, since human beings have limited cognitive resources to spread over tasks, attention spent on one task is necessarily an opportunity cost of reducing the attention available for others, see Kahneman (1973). This limited attention hypothesis suggests that the quality of stock analysis may deteriorate as the number of stocks users follow increases. StockTwits is the perfect platform to test this hypothesis by tracking the number of unique stocks and industries a user tweeted and calculating a *Sentiment* value for each user, firm in each month.

In chapter 4, my first finding is that social-media sentiment predicts positively and significantly future stock returns controlling a wide range of firm characteristics with the user and year-month fixed effects. The positive predictability of social-media sentiment decreases with the number of stocks and industries followed. These are important findings that not only does it confirm the validity of the limited attention hypothesis, the results also continue to hold when excluding users who declare themselves technical and short-term traders, who do not formulate trading ideas based on company fundamental news.

My regression result contributes to the prediction of social-media sentiment on a company's earnings positively and significantly, and such predictability decreases with the

investor's expanding stock and industry coverage, confirming the limited attention hypothesis that one can process a very restricted amount of information at any given time. My second finding is a trading strategy to generate a significant monthly excess return of 21.7 basis points and a trading alpha of 26.4 basis points per month. To do so, the users are divided into low- and high-stock-coverage group with averaged social media net sentiment. The strategy is to long a portfolio of stocks with zero and positive net sentiment, and shorts a portfolio of stocks with the negative net sentiment. The contribution of such policy implication in this study reveals that the individual investors are better off with low stock coverage due to limited attention. However, future research could be applied to other social-media platforms to find the result consistency.

In Chapter 5, I studied a long-standing problem that spans several decades in testing the CAPM model empirically where higher risks do not always deliver higher returns. Several academics suggest that investors are limited by short-sell constraint whereby over-priced stocks, as appeared in those high beta stocks, are prohibited from short-sell leading to a lower future return. Recent research explored how investor's disagreement in stock market affect asset's risk and future return.

To explore social-media disagreement against beta risk, I quantified and aggregated the user-level social-media disagreement from bottom-up to create the stock-level disagreement. The stocks were sorted into twenty portfolios according to their betas. The disagreement of stocks in each portfolios were averaged into a portfolio-level-disagreement. My third finding confirms that social-media investors disagree more on high-beta stocks, the portfolio-level social-media disagreement increases linearly with beta; and the shape of the portfolio returns appears to be concave. Moreover, the disagreement measure can further aggregate into stock-market level as a proxy for macroeconomic disagreement. The purpose of this aggregation is to see how it is related to other macroeconomic indices. My forth finding is that the aggregated social media disagreement can be determined and predicted by various macroeconomic indicators, i.e., C*onsumption-wealth ratios*, *Dividend Price ratio*, *default spread*, and *term spread*, to name a few. For instance, CBOE Volatility Index (VIX), also known as the fear index, predicts significantly higher aggregate social-media disagreement.

Having established the baseline relationships between social-media disagreement and returns. Hong and Sraer (2016)'s proposed to solve the high risk, low return problem by including short-sell constraint and relaxing a fundamental assumption in the CAPM model where

individual investors have homogeneous expectations. Two groups of optimist and pessimist analysts are formed who disagree on the fundamentals of a stock. Low-beta stocks, such as a utility company, are less prone to disagreement due to their stable fundamentals; whereas, high-beta stocks are more subject to scrutiny and disagreement such as the start-up companies, which are more likely to have fundamental uncertainty and firm-specific risk. Intuitively, when disagreement in both macro and firm-specific components are high, pessimistic investors are side-lined from short-selling high-beta stocks, leading to lower future returns.

Using the I/B/E/S analyst forecast dispersion with the sample period from 1984 to 2014, Hong and Sraer (2016) predicted that in low aggregate disagreement month Securities Market Line remains a positive linear relationship between risk and expected return according to CAPM; whereas the concavity only appears in high aggregate disagreement months. However, their hypothesis failed to hold for the new sample period of StockTwits dataset from 2010 to 2017. My fifth finding in the StockTwits dataset (2010 to 2017) produced a positive linear relationship upon graphical examination during the high aggregate disagreement months in forward future 9, 12, 18-month return. Such finding concurred with the result of replicating the analysis using the I/B/E/S analyst forecast dispersion under the same sample period as the StockTwits dataset. The inconsistency implies that the data sampling period matters which challenges the Hong and Sraer (2016)'s theory, since market microstructure has changed which allows more individual investor to short-sell, for example, through online trading app in recent years.

A regression analysis was used to examine whether aggregate social-media disagreement explains the concavity of the Securities Market Line. This test follows Fama and Macbeth (1973) style regression whereby in the first stage, I regress forward future return on beta and beta squared in the cross-section and collect the coefficients, analogous to fitting a non-linear curve as the empirical Securities Market Line. The second stage regresses the time series coefficients on various macroeconomic factors that could explain the coefficients. In particular, the squared coefficient is of interest here because it captures the concavity of the Securities Market Line. Although such a method is performed in various studies, it suffers from error-in-dependent-variable where the coefficients may not have control over other confounding factors, which I believe is one of the limitations of the analysis. My results suggest that aggregate social-media disagreement is robust in explaining the squared coefficient at forward future 9, 12, 18-month

value-weighted beta-sorted portfolio return. Results are less significant with equally-weighted beta-sorted portfolio return.

In summary, the study of non-institutional investors as a proxy to investor disagreement using the StockTwits dataset is a new approach in understanding the relationship between systematic risk and expected stock return. The findings are robust with different aggregate disagreement measures at different investment horizons in line with Hong and Sraer (2016) result.

Overall, future research in econometrics can improve the methods in dealing with multi-level data, while research direction can be extended to an international setting where investors in different countries might experience different cultures in investing, leading to different outcomes. The next question is whether we could explore a further understanding of the interaction between asset pricing and investor disagreement if we were able to put more social-media platforms under the spotlight.

I wish that my work contributes to the merits of a deeper research in behavioural finance, and also let the social-media communities see the benefits of making the data available for research.

7. References

Almazan, Andres, Keith C. Brown, Murray Carlson, and David Chapman. 2004. "Why Constrain Your Mutual Fund Manager?" *Journal of Financial Economics* 73 (2): 289–321.

Alves, Helena, Cristina Fernandes, and Mário Raposo. 2016. "Social Media Marketing: A Literature Review and Implications." *Psychology & Marketing* 33 (12): 1029–38.

Antweiler, Werner, and Murray Z. Frank. 2004. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *The Journal of Finance* 59 (3): 1259–94.

Baker, Malcolm P., Brendan Bradley, and Jeffrey Wurgler. 2010. "Benchmarks as Limits to Arbitrage: Understanding the Low Volatility Anomaly." Papers.Ssrn.com. Rochester, NY. March 1, 2010.

Baker, Malcolm, and Jeffrey Wurgler. 2006. "Investor Sentiment and the Cross-Section of Stock Returns." *The Journal of Finance* 61 (4): 1645–80.

Baker, Malcolm, Jeffrey Wurgler, and Yu Yuan. 2012. "Global, Local, and Contagious Investor Sentiment." *Journal of Financial Economics* 104 (2): 272–87.

Banz, Rolf W. 1981. "The Relationship between Return and Market Value of Common Stocks." *Journal of Financial Economics* 9 (1): 3–18.

Barber, B. M., and T. Odean. 2001. "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment." *The Quarterly Journal of Economics* 116 (1): 261–92.

Barberis, Nicholas, and Richard Thaler. 2002. "A Survey of Behavioral Finance." National Bureau of Economic Research Working Paper Series. September 19, 2002.

Bargeron, Leonce L., Kenneth Lehn, Sara B. Moeller, and Frederik P. Schlingemann. 2014. "Disagreement and the Informativeness of Stock Returns: The Case of Acquisition Announcements." *Journal of Corporate Finance* 25 (April): 155–72.

Barinov, Alexander. 2013. "Analyst Disagreement and Aggregate Volatility Risk." Papers.Ssrn.com. Rochester, NY. August 1, 2013.

Barron, Orie E., Donal Byard, Charles Kile, and Edward J. Riedl. 2002. “High-Technology Intangibles and Analysts’ Forecasts.” *Journal of Accounting Research* 40 (2): 289–312.

Bartov, Eli, Lucile Faurel, and Partha S. Mohanram. 2018. “Can Twitter Help Predict Firm-Level Earnings and Stock Returns?” *The Accounting Review* 93 (3): 25–57.

Basu, S. 1977. “Investment Performance Of Common Stocks In Relation To Their Price-Earnings Ratios: A Test Of The Efficient Market Hypothesis.” *The Journal of Finance* 32 (3): 663–82.

Berges, Angel, John J. McConnell, and Gary G. Schlarbaum. 1984. “The Turn-of-the-Year in Canada.” *The Journal of Finance* 39 (1): 185–92.

Bernard, Victor L., and Jacob K. Thomas. 1990. “Evidence That Stock Prices Do Not Fully Reflect the Implications of Current Earnings for Future Earnings.” *Journal of Accounting and Economics* 13 (4): 305–40.

Black, Fischer. 1972. “Capital Market Equilibrium with Restricted Borrowing.” *The Journal of Business* 45 (3): 444–55.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*.

Boni, Leslie, and Kent L. Womack. 2006. “Analysts, Industries, and Price Momentum.” *Journal of Financial and Quantitative Analysis* 41 (1): 85–109.

Bossaerts, Peter L., Elizabeth Bowman, Felix Fattinger, Harvey Huang, Carsten Murawski, Anirudh Suthakar, Shireen Tang, and Nitin Yadav. 2019. “Asset Pricing under Computational Complexity.” *SSRN Electronic Journal*.

Bradley, Daniel, Sinan Gokkaya, and Xi Liu. 2017. “Before an Analyst Becomes an Analyst: Does Industry Experience Matter?” *The Journal of Finance* 72 (2): 751–92.

Brandt, Michael W., Alon Brav, John R. Graham, and Alok Kumar. 2009. “The Idiosyncratic Volatility Puzzle: Time Trend or Speculative Episodes?” *Review of Financial Studies* 23 (2): 863–99.

Campbell, John Y. 1987. “Stock Returns and the Term Structure.” *Journal of Financial Economics* 18 (2): 373–99.

Carhart, Mark M. 1997. “On Persistence in Mutual Fund Performance.” *The Journal of Finance* 52 (1): 57–82.

Chan, Kalok, and Allaudeen Hameed. 2006. “Stock Price Synchronicity and Analyst Coverage in Emerging Markets.” *Journal of Financial Economics* 80 (1): 115–47.

Clement, Michael B. 1999. “Analyst Forecast Accuracy: Do Ability, Resources, and Portfolio Complexity Matter?” *Journal of Accounting and Economics* 27 (3): 285–303.

Clement, Michael B., Lisa Koonce, and Thomas J. Lopez. 2007. “The Roles of Task-Specific Forecasting Experience and Innate Ability in Understanding Analyst Forecasting Performance.” *Journal of Accounting and Economics* 44 (3): 378–98.

Clement, Michael B., and Senyo Y. Tse. 2005. “Financial Analyst Characteristics and Herding Behavior in Forecasting.” *The Journal of Finance* 60 (1): 307–41.

Cohen, Lauren, and Dong Lou. 2012. “Complicated Firms.” *Journal of Financial Economics* 104 (2): 383–400.

Cohen, Randolph B., Christopher Polk, and Tuomo Vuolteenaho. 2005. “Money Illusion in the Stock Market: The Modigliani-Cohn Hypothesis.” Www.Nber.org. January 10, 2005.

Cookson, J. Anthony, and Marina Niessner. 2019. “Why Don’t We Agree? Evidence from a Social Network of Investors.” *The Journal of Finance*, November.

Corwin, Shane A., and Jay F. Coughenour. 2008. “Limited Attention and the Allocation of Effort in Securities Trading.” *The Journal of Finance* 63 (6): 3031–67.

Cronqvist, Henrik, Tomislav Ladika, and Zacharias Sautner. n.d. “Limited Attention to Detail in Financial Markets.”

Cross, Frank. 1973. “The Behavior of Stock Prices on Fridays and Mondays.” *Financial Analysts Journal* 29 (6): 67–69.

Cukierman, Alex, and Paul Wachtel. 1979. “Differential Inflationary Expectations and the Variability of the Rate of Inflation: Theory and Evidence.” *American Economic Review* 69 (4): 595–609.

Custódio, Cláudia, Miguel A. Ferreira, and Pedro Matos. 2013. “Generalists versus Specialists: Lifetime Work Experience and Chief Executive Officer Pay.” *Journal of Financial Economics* 108 (2): 471–92.

Custódio, Cláudia, Miguel A. Ferreira, and Pedro Matos. 2019. “Do General Managerial Skills Spur Innovation?” *Management Science* 65 (2): 459–76.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. 2011. “In Search of Attention.” *The Journal of Finance* 66 (5): 1461–99.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. 2014. “The Sum of All FEARS Investor Sentiment and Asset Prices.” *Review of Financial Studies* 28 (1): 1–32.

De Giorgi, Enrico G., Thierry Post, and Atakan Yalçın. 2019. “A Concave Security Market Line.” *Journal of Banking & Finance* 106 (September): 65–81.

Dean, Geoff, and Peter Bell. 2012. “The Dark Side of Social Media : Review of Online Terrorism.” *Pakistan Journal of Criminology* 3 (4): 191–210.

Dechow, Patricia M, Richard G Sloan, and Amy P Sweeney. 1995. “Detecting Earnings Management.” *The Accounting Review* 70 (2): 193–225.

Dellavigna, Stefano, and Joshua M. Pollet. 2009. “Investor Inattention and Friday Earnings Announcements.” *The Journal of Finance* 64 (2): 709–49.

Diether, Karl B., Christopher J. Malloy, and Anna Scherbina. 2002. "Differences of Opinion and the Cross Section of Stock Returns." *The Journal of Finance* 57 (5): 2113–41.

Dische, Andreas. 2002. "Dispersion in Analyst Forecasts and the Profitability of Earnings Momentum Strategies." *European Financial Management* 8 (2): 211–28.

"Disclosing New Data to Our Archive of Information Operations." 2019. Twitter.com. 2019. https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.html.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2002. "Securities Lending, Shorting, and Pricing." *Journal of Financial Economics* 66 (2-3): 307–39.

Dunn, Kimberly A., and Brian W. Mayhew. 2004. "Audit Firm Industry Specialization and Client Disclosure Quality." *Review of Accounting Studies* 9 (1): 35–58.

Efthimion, Phillip, Scott Payne, and Nicholas Proferes. 2018. "Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots." *Article* 1 (2).

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 383.

Fama, Eugene F., and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–65.

Fama, Eugene F., and Kenneth R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

Fama, Eugene F., and James D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81 (3): 607–636.

Ferreira, Daniel, and Raaj Kumar Sah. 2012. "Who Gets to the Top? Generalists Versus Specialists in Managerial Organizations." *SSRN Electronic Journal* 43 (4): 577–601.

Figlewski, Stephen. 1981. “The Informational Effects of Restrictions on Short Sales: Some Empirical Evidence.” *The Journal of Financial and Quantitative Analysis* 16 (4): 463.

Frazzini, Andrea, and Lasse Heje Pedersen. 2014. “Betting against Beta.” *Journal of Financial Economics* 111 (1): 1–25.

Giannini, Robert, Paul Irvine, and Tao Shu. 2017. “Nonlocal Disadvantage: An Examination of Social Media Sentiment.” *The Review of Asset Pricing Studies* 8 (2): 293–336.

Gilson, Stuart C., Paul M. Healy, Christopher F. Noe, and Krishna G. Palepu. 2001. “Analyst Specialization and Conglomerate Stock Breakups.” *Journal of Accounting Research* 39 (3): 565–82.

Grinblatt, Mark, Seppo Ikäheimo, Matti Keloharju, and Samuli Knüpfer. 2016. “IQ and Mutual Fund Choice.” *Management Science* 62 (4): 924–44.

Grinblatt, Mark, Matti Keloharju, and Juhani Linnainmaa. 2011. “IQ and Stock Market Participation.” *The Journal of Finance* 66 (6): 2121–64.

Grossman, Sanford, and Joseph Stiglitz. 1980. “On the Impossibility of Informationally Efficient Markets.” *American Economic Review* 70 (3): 393–408.

Harford, Jarrad, Feng Jiang, Rong Wang, and Fei Xie. 2018. “Analyst Career Concerns, Effort Allocation, and Firms’ Information Environment.” *The Review of Financial Studies* 32 (6): 2179–2224.

Harris, Milton, and Artur Raviv. 1993. “Differences of Opinion Make a Horse Race.” *Review of Financial Studies* 6 (3): 473–506.

Hiraki, Takato, Ming Liu, and Xue Wang. 2015. “Country and Industry Concentration and the Performance of International Mutual Funds.” *Journal of Banking & Finance* 59 (October): 297–310.

Hirshleifer, David, Sonya S. Lim, and Siew Hong Teoh. 2011. “Limited Investor Attention and Stock Market Misreactions to Accounting Information.” *Review of Asset Pricing Studies* 1 (1): 35–73.

Hirshleifer, David, and Siew Hong Teoh. 2003. “Limited Attention, Information Disclosure, and Financial Reporting.” *Journal of Accounting and Economics* 36 (1–3): 337–86.

Hong, Harrison, and David A. Sraer. 2016. “Speculative Betas.” *The Journal of Finance* 71 (5): 2095–2144.

Hong, Harrison, and Jeremy C. Stein. 1999. “A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets.” *The Journal of Finance* 54 (6): 2143–84.

Ivkovic, Zoran, and Scott Weisbenner. 2005. “Local Does as Local Is: Information Content of the Geography of Individual Investors’ Common Stock Investments.” *The Journal of Finance* 60 (1): 267–306.

Jegadeesh, Narasimhan, and Sheridan Titman. 1993. “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.” *The Journal of Finance* 48 (1): 65–91.

Jensen, Michael C, Fischer Black, and Myron S Scholes. 2019. “The Capital Asset Pricing Model: Some Empirical Tests.” Ssrn.com. 2019.

Jung, Boochun, Kevin Jialin Sun, and Yanhua Sunny Yang. 2012. “Do Financial Analysts Add Value by Facilitating More Effective Monitoring of Firms’ Activities?” *Journal of Accounting, Auditing & Finance* 27 (1): 61–99.

Kahneman, Daniel. 1973. *Attention and Effort.* Englewood Cliffs, N.J., Prentice-Hall.

Kahneman, Daniel, and Amos Tversky. 1972. “Subjective Probability: A Judgment of Representativeness.” *Cognitive Psychology* 3 (3): 430–54.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47 (2): 263–91.

Kandel, Eugene, and Neil Pearson. 1995. “Differential Interpretation of Public Signals and Trade in Speculative Markets.” *Journal of Political Economy* 103 (4): 831–72.

Koski, Jennifer Lynch, and Jeffrey Pontiff, 1999, How are derivatives used? Evidence from the mutual fund industry, Journal of Finance 54, 791–816.

Kyriazidou, Ekaterini. 1997. “Estimation of a Panel Data Sample Selection Model.” *Econometrica* 65 (6): 1335.

Lai, Sandy, and Melvyn Teo. 2008. “Home-Biased Analysts in Emerging Markets.” *The Journal of Financial and Quantitative Analysis* 43 (3): 685–716.

Lamont, Owen A. 2002. “Macroeconomic Forecasts and Microeconomic Forecasters.” *Journal of Economic Behavior & Organization* 48 (3): 265–280.

Lamont, Owen A., and Richard H. Thaler. 2000. “Can the Market Add and Subtract? Mispricing in Tech Stock Carve-Outs.” *SSRN Electronic Journal*.

Lamont, Owen A, and Jeremy C Stein. 2004. “Aggregate Short Interest and Market Valuations.” *American Economic Review* 94 (2): 29–32.

Lee, Charles M.C., and Bhaskaran Swaminathan. 2000. “Price Momentum and Trading Volume.” *The Journal of Finance* 55 (5): 2017–69.

Lettau, Martin, and Sydney Ludvigson. 2001. “Consumption, Aggregate Wealth, and Expected Stock Returns.” *The Journal of Finance* 56 (3): 815–49.

Lintner, John. 1965. “Security Prices, Risk, And Maximal Gains From Diversification*.” *The Journal of Finance* 20 (4): 587–615.

Loughran, Tim, and Bill Mcdonald. 2011. “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” *The Journal of Finance* 66 (1): 35–65.

Malmendier, Ulrike, and Geoffrey Tate. 2005. “CEO Overconfidence and Corporate Investment.” *The Journal of Finance* 60 (6): 2661–2700.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers. 2004. “Disagreement about Inflation Expectations.” Www.Nber.org. June 23, 2004.

Mezias, John M., and Stephen J. Mezias. 2000. “Resource Partitioning, the Founding of Specialist Firms, and Innovation: The American Feature Film Industry, 1912–1929.” *Organization Science* 11 (3): 306–22.

Miller, Edward M. 1977. “Risk, Uncertainty, and Divergence of Opinion.” *The Journal of Finance* 32 (4): 1151.

Newey, Whitney K., and Kenneth D. West. 1987. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica* 55 (3): 703.

Nickell, Stephen. 1981. “Biases in Dynamic Models with Fixed Effects.” *Econometrica* 49 (6): 1417–26.

Panousi, Vasia, and Dimitris Papanikolaou. 2012. “Investment, Idiosyncratic Risk, and Ownership.” *The Journal of Finance* 67 (3): 1113–48.

Pastor, Lubos, Robert F. Stambaugh, and Lucian A. Taylor. 2017. “Do Funds Make More When They Trade More?” *The Journal of Finance* 72 (4): 1483–1528.

Peng, Lin, and Wei Xiong. 2006. “Investor Attention, Overconfidence and Category Learning.” *Journal of Financial Economics* 80 (3): 563–602.

Piotroski, Joseph D., and Darren T. Roulstone. 2004. “The Influence of Analysts, Institutional Investors, and Insiders on the Incorporation of Market, Industry, and Firm-Specific Information into Stock Prices.” *The Accounting Review* 79 (4): 1119–51.

Renault, Thomas. 2017. “Intraday Online Investor Sentiment and Return Patterns in the U.S. Stock Market.” *Journal of Banking & Finance* 84 (November): 25–40.

Ritter, Jay. 2003. “Behavioral Finance.” *Pacific-Basin Finance Journal* 11 (4): 429–437.

Roof, Katie. 2016. “StockTwits Raises Funding, Gets New CEO.” TechCrunch. TechCrunch. July 6, 2016.

Sadorsky, Perry. 2003. “The Macroeconomic Determinants of Technology Stock Price Volatility.” *Review of Financial Economics* 12 (2): 191–205.

Seasholes, Mark S., and Ning Zhu. 2010. “Individual Investors and Local Bias.” *The Journal of Finance* 65 (5): 1987–2010.

Serafeim, George, and Aaron Yoon. 2021. “Stock Price Reactions to ESG News: The Role of ESG Ratings and Disagreement.” Papers.ssrn.com. Rochester, NY. January 13, 2021.

Sharpe, William F. 1964. “Capital Asset Prices: A Theory Of Market Equilibrium Under Conditions of Risk*.” *The Journal of Finance* 19 (3): 425–42.

Sonney, Frédéric. 2007. “Financial Analysts’ Performance: Sector Versus Country Specialization.” *Review of Financial Studies* 22 (5): 2087–2131.

Statman, Meir. 1987. “How Many Stocks Make a Diversified Portfolio?” *The Journal of Financial and Quantitative Analysis* 22 (3): 353.

Statman, Meir. 2004. “The Diversification Puzzle.” *Financial Analysts Journal* 60 (4): 44–53.

Swaminathan, Anand. 2001. “Resource Partitioning and the Evolution of Specialist Organizations: The Role of Location and Identity in the U.S. Wine Industry.” *Academy of Management Journal* 44 (6): 1169–85.

Tetlock, Paul C. 2007. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” *The Journal of Finance* 62 (3): 1139–68.

Thompson, Samuel B. 2011. “Simple Formulas for Standard Errors That Cluster by Both Firm and Time.” *Journal of Financial Economics* 99 (1): 1–10.

Tversky, Amos, and Daniel Kahneman. 1973. “Availability: A Heuristic for Judging Frequency and Probability.” *Cognitive Psychology* 5 (2): 207–32.

Wang, Diamond. 2019. “The Impact of Fund and Managerial Heterogeneities on Mutual Fund Performance.” Papers.ssrn.com. Rochester, NY. March 30, 2019.

Yu, Jialin. 2011. “Disagreement and Return Predictability of Stock Portfolios.” *Journal of Financial Economics* 99 (1): 162–83.

Zhang, Chu. 2010. “A Reexamination of the Causes of Time-Varying Stock Return Volatilities.” *Journal of Financial and Quantitative Analysis* 45 (3): 663–84.

Zhang, X. Frank. 2006. “Information Uncertainty and Stock Returns.” *The Journal of Finance* 61 (1): 105–37.