University of BRISTOL

Who Needs Due-Diligence?

An investigation into the heterogeneous impacts of Reddit sentiments upon asset trading volumes using quantile regression

Student Name: Nishal Dave

Student Number: 2079639

Word Count: 12,475

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science by advanced study in Economics in the Faculty of Economics

School of Economics

September 2021

ABSTRACT

Discussion about markets and stocks have proliferated nearly every platform of social media since the dawn of internet conversation. Research has explored how online conversations and the networks they develop influence the trading behaviour of individuals and the subsequent stock market characteristics. This study homes in on a subset of the wider Reddit community called WallStreetBets to evaluate the heterogenous impact of collective behaviour upon the trading volumes of stocks. By observing both the volume of conversation and the underlying sentiment, a quantile regression approach is then used to regress trading volume upon the these variables to recover the quantile level coefficients. Findings in this study show significant causal relationships between the growth in the volume of conversation and trading volume, similar results are also found for the underlying sentiment upon trading volume, and both remain robust to restrictions in sample and changes in specification. Overall, findings indicate meaningful variation in trading volumes explained by Reddit activity.

DEDICATION AND ACKNOWLEDGEMENTS

I would like to thank my supervisor, Sukjin (Vincent) Han, for his inspiration and guidance during this ambitious dissertation. I would also like to extend my appreciation to my parents and brother for their continued and unconditional support during my academic career and beyond.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

SIGNED: NISHAL DAVE

DATE: 27/09/2021

Table of Contents

1 INTRODUCTION 5

2 LITERATURE REVIEW 7

3 DATA 11

4 METHODOLOGY 13

PARAMETER ESTIMATION 16

5 ESTIMATION RESULTS 16

FULL SAMPLE 16

TOP 25% SUB-SAMPLE 17

MIDDLE 50% SUB-SAMPLE 18

BOTTOM 25% SUB-SAMPLE 18

DISCUSSION 19

6 ROBUSTNESS CHECKS 20

STOCKS WITH GREATER THAN 200 SUBMISSIONS 20

EXCLUDING OUTLIERS 20

ADDITION OF INTERACTION TERM 21

7 CONCLUSION 21

BIBLIOGRAPHY 24

APPENDIX 27

A.1: VARIABLES 27

A.2: TABLES 27

1 INTRODUCTION

The last decade has seen a substantial surge in social media usage, with around half of the global population actively using one or more platforms. The impact of these online interactions are becoming more and more apparent in the financial markets.

Advances in technology have levelled the playing field for both institutional investors and retail investors alike, with the ability to buy and sell securities using mobile devices combined with exposure to large scale discussion through social media. There are now fewer barriers to prevent the average individual from participating in the trading of securities.

Among these platforms exists Reddit, a platform which was established in 2005 and serves as a hub for discussion within individual communities called subreddits. With over 130,000 active subreddits on Reddit, communities exist across nearly every imaginable topic which attracts around 430 million monthly users as of July 2021.[1] Reddit supports various forms of submissions, including but not limited to pictures, video, microblogs, and external links on which other users can comments and interact amongst one another. The difference between Reddit and other popular social media is the complete anonymity of users, unlike Facebook and Twitter. This is favourable to those who have concerns over their privacy.

In 2012 a new subreddit named WallStreetBets was conceived. This community consisted mainly of discussions about company stocks and particularly options. The nature of the subreddit differed to other investing communities through its colourful use of language and terminology which attracted a crowd of younger individuals claiming to be self-described degenerates who have an excessive appetite for risk. Examples of submissions in this subreddit include screenshots of gains and losses made on highly leveraged positions or threads of discussions surrounding popular stocks at a given time. Stocks of interest are not limited to any specific industry, however there is a tendency towards newer and less established companies often within the technology sector including, biotech, pharmaceuticals, and software.

WallStreetBets gained significant attention in 2021 due to its users collectively inflating the stock price of GameStop following a period of posts made by a single user. This caused an increase in price of around 2,200% during January. Aside from the profit made by many of the investors, several institutional investors had large short positions against GameStop which totalled an approximate loss of six billion dollars for institutional investors

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

and the eventual closure of a hedge fund due to substantial losses.

This increase in popularity led to a surge in the number of subscribers to the subreddit, which doubled in a week to five million users during the second half of January 2021, as of recently the subscriber count stands at ten and a half million. Using a conservative estimate of $500 per subscriber brings the purchasing power of WallStreetBets at $5.25 billion, which is more than enough for a coordinated effort to influence market outcomes.

These events followed a similar pattern of behaviour described by Sabherwal, Sarkar, and Zhang (2011), a surge in the stock price that was driven by hype and consensus amongst readers of the forum, which stimulated trading activity. Semenova and Winkler (2021) also find a significant impact of Reddit's Wallstreetbets conversation on trading volume across a wide sample of stocks, they use the concepts of contagion and consensus as avenues to measure the spread of discussion over time and the extent to which users agree on a direction of price movement respectively.

Existing literature and this series of events has served as an inspiration for this paper, which aims to model the heterogenous impacts of WallStreetBets collective behaviour upon trading volume. Trading volume of a particular security can act as an indicator for the amount of interest it is receiving, as it is simply defined by the volume of buying and selling taking place. Establishing whether trading volumes are influenced by behaviour on Reddit could prove to be a valuable tool in determining stocks that are increasing in popularity.

This study uses two measures of collective behaviour which take the form of sentiments from text submissions and submission volumes drawn from WallStreetBets forum. These variables are then used to model variation in the trading volume of a panel of discussed assets using a quantile regression approach. Quantile regression enables modelling of potential heterogeneity in impacts across the conditional quantile distribution of trading volume, this is of particular importance as trading volume generally has a fat-tailed distribution so the OLS effect may not adequately capture the true impact of collective behaviour across the entire distribution of trading volumes.

The expectations from this study follows that firm which have a lower value are likely to be more affected by collective behaviour in comparison to firms which have a higher value. Additionally, we expect to see impacts to be pronounced in the higher quantiles of trading volume.

It would be expected that the firms in the lowest valuation sample are likely to experience most variation in trading volumes induced by discussion on WallStreetBets. This is because smaller firms are subjected to less sources of variation such as fewer institutional investors and less discussion across various media. Therefore, collective behaviour on Reddit

would likely make up a larger proportion of these firms' trading volume compared to larger firms.

As a general pattern, the larger the firm, the larger the trading volume is likely to be. Although this is not a rule and there are likely to be many exceptions to this case. For example, smaller firms that experience an up-shoot in popularity may occupy the upper quantiles of the conditional quantile distribution of trading volume. Therefore it would be expected to see much of the variation occuring towards the top of the conditional quantile distribution.

This will be supported by a subsampling approach accounting for market capitalisation to control for different firm sizes, as the whole sample may not paint the most representative picture.

The findings in the study support our expectations as there is a pronounced impact of submission volume and submission sentiment upon trading volumes across quantiles. With an increasing effect towards the top of the distribution of trading volumes, there is a noticeable heterogeneous effect which is stronger for smaller firms in the sample.

The paper is organised as follows, the next section presents the literature review, followed by section 3 which describes the data and collection process. Section 4 outlines the methodology and model specification whilst section 5 examines the results of the regression and discussion followed by an evaluation of robustness in section 6, finally Section 7 presents the concluding remarks.

2 LITERATURE REVIEW

The following literature capture conversation from social platforms. With the advancement of artificial intelligence and natural language processing technologies, sentiments can be derived from conversations and used for both inferential and predictive analyses.

Gruhl, Guha, Kumar, Novak, and Tomkins (2005) evidence how the volume of "online chatter" shows explanatory power through strong cross-correlation between blog mentions and sales rank in the context of Amazon books, this analysis sets a precedent in modelling online behaviour.

Bollen, Mao, and Zeng (2011) present a relationship between large scale Twitter sentiments and the daily change in the price of the Dow Jones Industrial Average index (DJIA) using Granger Causality analysis. The results show significant granger causality between the calm mood and a change in the DJIA price, whilst all other moods have no significant effect. One may expect this result as higher levels of calmness may imply more attention to judgement.

A potential pitfall in the GPOMS methodology is that the set of tweets considered are only those that include the quotes "I am feeling" and "I don't feel", Martinez-Camara, m. Teresa Martin-Baldivia, l. Alfonso Urena-Lopez, and

Montejo-Raez (2012). This may disregard information from Tweets that may not explicitly state how the author feels but captures valuable sentiment through tone of text and synonymous language. Sprenger, Tumasjan, g. Sandner, and Welpe (2013) find similar relationships between tweet sentiment and returns, tweet volume and trading volume. In contrast to the hypothesis of bullish sentiment impacting returns, a reverse relationship is observed. Abnormal returns are likely to increase bullish sentiment across Twitter, highlighting an endogenous relationship between sentiments and market features.

Deng, Huang, Sinha, and Zhao (2018) observe a significant relationship between DJIA returns and Stocktwits sentiments on an hourly timeframe using a Vector Autoregression (VAR) methodology, which accounts for the endogeneity previously noted by Sprenger et al. (2013). Findings show a significant relationship at the hourly level, which previous literature did not address. This study uncovers the larger role of negative sentiment in explaining the behaviour of noise traders. There are also asymmetries in the response of negative sentiment against positive sentiment, which are intriguing and well documented with roots in prospect theory, Kahneman and Tversky (1979). Investors apply weight asymmetrically to gains and losses, where a loss is perceived as worse compared to an equivalent gain, this could be explanatory as to why investors are more responsive to negative sentiment compared to being reluctant despite the presence of positive sentiment.

Ganesh and Iyer (2021) also adopt the VAR approach in the analysis of firm-initiated tweets upon stock return and trading volume of stocks in the DJIA, however they focus on frequency of the tweets rather than content of the message. Compared to previous analyses mentioned, trading volume seems to be more responsive to the frequency of tweets. Considering the prominence of the DJIA and its contained firms (many of which have a substantial twitter following) it would be expected that more frequent tweets would garner and maintain relevance, this is likely to increase trading activity of these stocks. The results also present an endogenous link between the frequency of tweets and trading volume.

Michalak (2020) explores the relationship between Twitter sentiments and trading volumes of Facebook, Apple and Amazon using a Naive Bayes Classifier for Tweets over a two-year period. She finds significant results for negative sentiments upon trading volumes.

Our study differs from the literature mentioned above in a handful of ways. Firstly, none of the literature so far have utilised the same text sentiment algorithm used in this paper. They are also geared towards the modelling of stock returns rather than volume. However this is still important to review, as common interpretations such as asymmetry in investor behaviour are encountered in our findings.

There have been fewer attempts at non-linear modelling in the literature so far. Ni, Wang, and Xue (2015) find significant effects of sentiments upon returns using a Panel Quantile Regression approach pioneered by Koenker and Bassett (1978). This method aims to capture the heterogeneity across quantiles in sentiment on returns in the Shanghai A-share stock market. In comparison to the previous literature mentioned, sentiment is captured by the number of open accounts and the turnover rate as they are thought to proxy investor confidence and liquidity respectively, Baker and Stein (2004). Comparatively Ma and Xiao (2018) find significant and negative coefficients only at the lower end of the distribution of excess returns of the S&P 500 when considering the impact of the BW and the HJTZ sentiment indices which are developed in Baker and Wurgler (2006) and Huang, Jiang, Tu, and Zhuou (2015) respectively. Both analyses show effects beyond the mean, which suggest there are clear non-linearities which had not been previously identified.

Swamy and Dharani (2020) take a quantile regression approach to the estimation of returns in the Indian stock market upon Google search volume. Findings show that the volume of Google searches relating to a given firm are both positive and significant in explaining returns. The findings here are consistent with similar analyses carried out using the GSVI in the Russell 3000 and S&P 500 in the mid to late noughties by Da, Engelberg, and Gao (2011) and Joseph, Wintoki, and Zhang (2011) respectively. Vlastakis and N.Markellos (2012) use the GSVI (Google Search Volume Index) to measure the contemporaneous impact upon trading volume of firms in the DJIA using OLS. They find that most of the firm level coefficients are significant individually as well as the coefficient for the pooled sample.

There are some noteworthy results that have been found in the quantile regression approaches, however the literature of social media sentiments upon financial metrics at the quantile level remains sparse and based on the linear method outcomes and it may be a fruitful exercise to delve deeper. This aligns with the aims of our paper to explore the quantile level outcomes using different measures of collective behaviour in contrast to the existing literature.

Uninformed investors generally trade on trends, emotion, and other non-fundamentals, they are often labelled as 'noise-traders', Brown (1999). These individuals are largely driven by sentiment, he remarks that higher volatility introduces higher trading volume. This is supported by Black (1986) who claims that in the absence of noise, trading of assets would be an uncommon practice. This is based on the idea that trading requires two parties who have conflicting information, if both parties proceed then one party is acting on wrong information. Proceeding with the trade this implies that they

have acted on sentiment or instinct, and enough of this behaviour creates noise, and increased trading frequency.

In line with the above Harris and Raviv (1993) concludes that trading is initiated by conflicting opinions between traders regarding the value of a stock, elevated disagreement results in more trading activity. Antweiler and Frank (2005) evaluate the impact of forum messages and their impact on the 45 companies within the DJIA using a unidimensional signal of bullishness or bearishness. Findings show that disagreement among the messages increases trading volume on the same day, adding support to Harris and Raviv (1993). On the other hand, next day trade volume decreases in the face of previous day disagreement, whilst agreement among messages has a positive impact for trading volumes in future periods.

Additionally, the volume of messages is also found to be a significant determinant of trade volume. Comparably, Li, Dalen, and van Rees (2018) performs both an inter and intraday analysis using sentiments drawn from Tweets and finds that agreement tends to reduce trading volume in both inter and intraday settings, however the intraday effects contrast those found by Antweiler and Frank (2005). Another observation is the role of volatility in determining trading volumes, previous day volatility increases the volume of tweets, which in turn generates uncertainty about future prices and initiates further discussion, if uncertainty creates more disagreement, this is likely to lead to even further increases in trading volume in support of both Brown (1999) and Michalak (2020) earlier findings.

Tetlock (2007) finds that either end of unusually high or low pessimism predicts an increase in following day trading volume this using the GI (General Inquirer) sentiment analysis tool and a VAR approach. He analyses the daily variation in moods related to DJIA and NYSE trading volumes by controlling for pessimism. This makes sense as polarisation in sentiments in this case would imply agreement in beliefs, which are again akin to the findings in Antweiler and Frank (2005) on the other hand these findings contradict those found by Li et al. (2018).

The literature so far has made leaps and bounds in evaluating the impact of various sources of sentiment upon returns and trading volume, and there is ample evidence in support of a significant relationship between the two. With the progression of time, sentiments have been sourced from an increasing number of media which now include Reddit, and what was once a group of individuals known for their incessant appetite for risk and crude humour has evolved into a community with significant purchasing power and the ability to turn tides in the market.

Additionally, there have been a variety of methods use to model these relationships, a common one being the traditional VAR method that accounts for endogenous feedback, although this relies

upon a linear assumption which is limiting in the scope of sentiment impacts. Alternatively, the more recent applications of quantile regression have demonstrated heterogenous relationships across conditional quantiles between these variables, which is indicative of non-linearity. However, a true attempt at modelling the heterogenous effects of social media sentiments from remains absent, which paves the way for this dissertation to estimate the effects of Reddit sentiments at the quantile level on asset trading volumes.

3 DATA

This study examines 110 stocks over a period of a year between 06/2020 to 06/2021 in an unbalanced longitudinal panel using weekly intervals.[2]

We select the sample based on the most mentioned stock tickers on the Wallstreetbets forum. We then began by constructing a Python script to aggregate the number of submissions titles containing the ticker symbol of each stock in the year period of 06/2020 to 06/2021, all requests were queried via the Pushshift API. The keyword used is the ticker symbol identified by the dollar symbol '$' followed by a string of characters representing a firms' stock. This approach is sufficient as it follows the communication style of users in the WallStreetBets community and therefore recovers the appropriate submissions. This consisted of every stock traded across the three major stock exchanges in the United States which are NASDAQ, New York Stock Exchange, and the NYSE American, which form a total of 7,830 listed stocks. Most stocks across the exchanges are not mentioned at all on the forum, however for completeness they were included in the search.

To gain a representative opinion for each stock, we imposed a lower limit on the number of mentions for each ticker. Therefore those with less than 100 mentions are excluded from the sample outright, which eliminates 7428 stocks. Additionally, many ticker symbols take on symbols which represent words such as '$IS' or '$BIG', they have many occurrences although they are out of context of the underlying stock and are excluded, this is because they are simply sources of noise. Due to limitations of the Pushshift API, sorting through those which were representative of the stock against the irrelevant tickers was a manual process which required searching through top Wallstreetbets discussions for obvious mentions of the stock or whether the term only contributed to overall noise as opposed to underlying stock discussion.

This brought the resulting sub-sample down to 110 viable stocks with adequate discussion to use in this analysis. The source of imbalance in the panel structure is due nine firms in the sample who have had their initial public offering during the term of study and therefore public stock data is not available for the entire 52

[2]See table 10 in Appendix A.2 for a complete list of firms

weeks being observed.

Financial data was queried using the AlphaVantage API in Python with additional transformation performed in R.[3] Raw data was obtained at the daily level for price and trading volume by number of shares.

To evaluate mood and feelings from these submissions, the body of the submission rather than title was used as the source of sentiment. This is because it contains more information which can produce a more representative value for sentiment.

To quantify the sentiment, we have used VADER (Valence Aware Dictionary and sEntiment Reasoner), which is a natural language processing tool which has been designed with social media analysis in mind, Hutto and Gilbert (2014). VADER takes string data from the body of the submission as an input and returns a number between -1 and 1 as the output. This score ranges from extremely negative to extremely positive respectively using a predefined lexicon to determine the score. This compound score serves as a uni-dimensional measure of sentiment that can be used for analysis.

The lexicon is populated by a list of words and an accompanying score which describes the semantic orientation of the word. The semantic orientation describes whether words are positive, neutral, or negative and is determined by humans using a manual procedure.

[3]https://www.alphavantage.co/

Sentiment analysis has been widely studied and there exist many tools and approaches to extrapolate feelings from a body of text, therefore VADER is one of many available packages. A similar contender is LIWC (Linguistic Inquiry and Word Count) which is suitable for analysing social media microblog style texts, Pennebaker, Booth, Boyd, and Francis (2015). In comparison to VADER, LIWC does not have the power to differentiate between the intensity of a message, for example the words “good” and “amazing” are both positive however the latter conveys a stronger feeling, LIWC only recognises that both words are positive and therefore they are given the same semantic orientation.

Adjectives are crucial in determining the tone of a message and this is where VADER comes out on top in its ability to incorporate the tone of the word in addition to its polarity. The general inquirer is one of the oldest and most popular tools, whilst it has an expansive lexicon, it too suffers from the same inability to distinguish the intensity of key words, Philip, Robert, Zvi, and Daniel (1962).

VADER acknowledges this issue and accounts for the tone of text which makes it a favourable option. The other advantages of VADER are reflected in its simplicity yet accuracy, this means significantly quicker computation in comparison to more common methods such as the Support Vector Machine or Naïve

Bayes approaches. VADER is also an open source and free to use tool as well as being very easy to integrate into the existing Pandas data frame, so implementation involves a few lines of code. In addition to this, as VADER is lexicon based rather than a black-box model, it is easy to directly observe the rules use to compute sentiments and even modify the lexicon if necessary.

Considering these factors, there is a compelling argument towards using VADER and therefore we have proceeded with it as our choice of sentiment analysis tool for this study.

4 METHODOLOGY

For the estimation, we have used a fixed effects mean model and a quantile regression approach introduced by (Koenker & Basset, 1978). Quantile regression provides flexibility in estimating heterogenous and distributional impacts across quantiles. By comparison Ordinary Least Squares (OLS) captures a limited amount of information as it only models the mean effect. Quantile regression can be informative to explain potential non-linearities in the data. This is achieved by recovering estimates for each exogenous regressor on each quantile, where each coefficient is now a function of the quantile probability τ.

In the context of this analysis, the presence of firm specific characteristics that correlate with both trading volume and the collective behaviour may be unobservable or have practical limitations to acquire, which requires the use of panel data methods which estimate this unobservable heterogeneity.

(Machado & Silva, 2019) have developed the Method of Moments Quantile Regression with fixed effects (MMQR) which takes a parsimonious approach to estimating the quantile effect in panel data, by allowing the individual heterogeneity to influence the whole distribution as opposed to just shifting location. Additionally the MMQR model manages the omitted variable bias in the same way that a regular fixed-effects model does as any omitted variable that remains constant over time is absorbed by the fixed effect which eliminates the unobserved heterogeneity.

The Penalised Fixed Effects model introduced by Koenker (2004) also addresses this challenge by using a shrinkage parameter λ as there are a large number of individual fixed effects. The complications arise in determining a suitable value for the shrinkage parameter, in the predictive landscape a cross-validation approach would be sufficient (James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013). Due to this, we use the MMQR model mentioned previously.

Our objective is to estimate the conditional quantiles of firm trading volume V_{it} conditional upon a five-dimensional set of exogenous regressors. The conditional distribution of V_{it} belongs to the location-

scale family of the following form:

$$V_{it} = \alpha + \mathbf{X}'_{it}\beta + \sigma(\delta_i + \mathbf{Z}'_{it}\gamma)U_{it},$$

$$U_{it} \sim \text{Uniform}(0, 1).$$

The vector $\mathbf{X}'_{it}$ contains the following regressors outlined below:[4]

$$S_{it-1} = E[s_{id}|d \in t-1],$$

S_{it-1} measures the previous week average sentiment value determined by the VADER algorithm, which is continuous between -1 and +1, this translates to how positive or negative an individual is about a stock. Semenova and Winkler (2021) also use this measure, although defined much differently as their variable is constructed based upon a series of influences which would lead an individual to make a buy or sell decision taking on -1 or +1 implying bearish or bullish which is modelled by a quantal response function. So although variable definitions vary between each other, the essence of the underlying information remains the same to the extent that they capture an individuals sentiment in regard to a given stock.

Wang, Yu, and Shen (2021) also use a lagged value of sentiment justified by the common notion that present values are influenced by past behaviour, providing additional confidence towards using a week-long lag.

$$\Delta A_{it} = \ln\Big(\sum a_{id}|d \in t\Big) - \ln\Big(\sum a_{id}|d \in t-1\Big),$$

ΔA_{it} measures the change in the number of submissions from week to week, this measure is also inspired by the model developed in Semenova and Winkler (2021) who use ΔA_{it} as a measure of social contagion. Social contagion is the concept of ideas spreading from one user to another user to generate more discussion, although they use the number of submissions by unique individuals. We employ a simpler definition by using the absolute number of new submissions. This is done in part due to sample size, using a smaller sample of firms means that each observation has more marginal importance, to preserve an adequate sample size we retain all submissions. In spite of this, absolute discussion volume should remain an important measure of interest, because enthusiasm around a given stock is still captured in the weekly variation.

$$V_{it-1} = \ln\big(E\big[p_{id} \cdot v_{id}|d \in t-1\big]\big),$$

including lagged trading volume V_{it-1} in the model controls for the autoregressive relationship and isolates any past variation. Due to the high frequency nature of trading volume a single week-long lag is adequate, additional lags are unlikely to provide further value, Wang et al. (2021).

$$\sigma_{it-1} = E\big(\sqrt{Var(r_{id}|d \in t-1)}\,\big),$$

$$Cov^R_{it} = cov(R_{it}, R_{it-1}),$$

the remaining variables control for past volatility σ_{it-1} and the first order autocovariance of returns Cov^R_{it}. Both of these variables are found to be significant in es-

[4]See Appendix A.1 for detailed variable definitions

timating trading volumes by Lo and Wang (2000) and Kumar (2006).

Due to the potential for endogeneity, we make a reasonable assumption that the regressors can only depend on lagged terms, this implies that the lags of sentiment, trading volume, and average volatility are not dependent on the present value of trading volume. By doing this the issue of simultaneity is managed. However, as trading volume has no impact on the autocovariance of returns, we maintain a contemporaneous term for this variable. Submissions volume is also used as weekly growth rather than a lag, however as the change from last week to the present week cannot be influenced by the present week's trading volume this excludes it from being endogenous with trading volume.

Additionally, stock returns have been the main focal point of previous studies in this field, however in the context of Reddit sentiments Semenova and Winkler (2021) find a very small impact on returns, despite using a richer data-set and more precise measures of sentiment.

This could be in line with the theory of Efficient Market Hypothesis, Fama (1970). the information reflected in retail sentiment is information that would be captured in stock returns, assuming a retail investor does not have access to privately held information, rather they formulate judgement based on information that is in the open domain. Alternatively, Reddit sentiments may not be influential enough to explain any significant variation in the stock market beyond creating noise, which Semenova and Winkler (2021) simulate in their paper.

For these reasons, there is no expectation of significance of WallStreetBets sentiments to influence stock returns.

The parameters of interest are,

$$(\alpha_i, \beta', \delta_i, \gamma')',$$

with (α_i, δ_i) representing the individual fixed effects.

$\mathbf{Z}_{it}$ is a five-dimensional vector of transformations of $\mathbf{X}_{it}$ and $\sigma(\cdot)$ is a known function.

The purpose of $\mathbf{Z}_{it}$ and the scale function $\sigma(\cdot)$ is to allow flexibility into the model. For this application the standard quantile regression assumes $\mathbf{X}_{it} = \mathbf{Z}_{it}$ and $\sigma(\cdot)$ is the identity function. As such, the location-scale form can be re-expressed as,

$$V_{it} = \alpha_i + \mathbf{X}'_{it}\beta + (\delta_i + \mathbf{X}'_{it}\gamma)U_{it}$$

The τ'th conditional quantile takes on the form:

$$\begin{aligned} Q_Y(\tau|X_{it}) &= \alpha_i + \mathbf{X}'_{it}\beta + (\delta_i + \mathbf{X}'_{it}\gamma)q(\tau) \\ &= \alpha_i + \delta_i q(\tau) + \mathbf{X}'_{it}(\beta + \gamma q(\tau)) \end{aligned}$$

where the term $(\alpha_i + \delta_i q(\tau))$ represents the time-invariant individual fixed effect for each stock i at quantile τ. More specifically, α_i represents the average fixed effect for stock i whilst δ_i controls the quantile level fixed effect.

PARAMETER ESTIMATION

To obtain $\hat{\beta}$ run the following least squares regression:

$$\left(V_{it} - \sum_t \frac{V_{it}}{T}\right) = \hat{\beta}\left(X_{it} - \sum_t \frac{X_{it}}{T}\right),$$

using the above, estimate $\hat{\alpha}$ and recover the residuals,

$$\hat{\alpha}_i = \frac{1}{T}\sum_t (V_{it} - X'_{it}\hat{\beta}), \quad \hat{R}_{it} = V_{it} - \hat{\alpha}_i - X'_{it}\hat{\beta}.$$

Use these residuals to obtain $\hat{\gamma}$ by the following least squares regression,

$$\left(|\hat{R}_{it}| - \sum_t \frac{|\hat{R}_{it}|}{T}\right) = \hat{\gamma}\left(X_{it} - \sum_t \frac{X_{it}}{T}\right),$$

using $\hat{\gamma}$, it is possible to estimate $\hat{\delta}_i$,

$$\hat{\delta}_i = \frac{1}{T}\sum_t \left(|\hat{R}_{it}| - X'_{it}\hat{\gamma}\right).$$

$q(\tau)$ can be estimated by choosing the $\hat{q}$ that minimises the problem below,

$$\min_q \sum_i \sum_t \rho_\tau\big(\hat{R}_{it} - (\hat{\gamma}_i + X'_{it}\hat{\gamma})q\big).$$

Where $\rho_\tau(\cdot)$ is the check function expressed below:

$$\rho_\tau(\Theta) = (\tau - 1)\Theta I\{\Theta \leq 0\} + \tau\Theta I\{\Theta > 0\}.$$

Before proceeding with any analysis, it is important to pre-test the data to ensure that it is absent of any unit roots that cause non-stationarity. The IPS, Im, Pesaran, and Shin (2003) unit root test is an appropriate test for unbalanced panel data, where the null states that all panels contain unit roots whilst the alternative states that one or more panels do not contain a unit root. Lag selection is determined by minimising the AIC. Appendix A.2, table 2 displays the results of the test on the variables of interest and rejects the null at all significance levels for all variables.

5 ESTIMATION RESULTS

The findings are organised as follows, the first set of results include the entire sample size of 110 stocks. The following three sets of results sub-sample the data by market capitalisation of the stocks, which are sorted by top 25%, middle 50% and bottom 25%. This is to discern the identified effects based on the value of the firm, as the value of a firm does not necessarily determine its trading volume. In addition to the OLS coefficients the results include each consecutive 0.2 quantile as well as the bottom 0.1, median 0.5 and top 0.9 quantiles of the conditional quantile distribution for trading volume. All regression results are presented in Appendix A.2.

FULL SAMPLE

The findings for the full sample experiment in table 3 show a positive and statistically significant impact of lagged sentiment upon trading volume. As the quantile probability increases, the impact of lagged sentiment upon trading volume also increases. A unit increase in lagged sentiment increases trading volume by 0.73% to 1.56% depending on the quantile of interest. At the mean level, a unit in-

[5] A unit pertains to a movement by 0.1 degree of VADER's sentiment compound value

crease in lagged sentiment increases trading volume by 1.1%.[5] The growth in submissions volume also exhibits a positive and statistically significant impact upon trading volume. Similarly to lagged sentiment, the impact of submissions growth upon trading volume increases as quantile probability increases. A 1% increase in submissions growth increases trading volume in the range from 0.096% to 0.28% depending on the quantile of interest. At the mean level, trading volumes increases by 0.18% for a 1% increase in submissions growth. This positive relationship is also observed by Antweiler and Frank (2005) who find that next period trading volume increases as the volume of messages or discussion increase.

The quantile impact shows a significant deviation from the mean, this exemplifies the importance of observing the heterogeneous impact.

Comparing mean outcomes, the results fall in line with Semenova and Winkler (2021) who conducted a study under a similar specification using an OLS approach, they show a positive and significant impact of both user sentiments and submissions growth upon trading volume.

Whilst it is not a key variable of interest, the impact of previous week's volatility is statistically and negative towards the bottom of the trading volume distribution. This provides interesting insight into the impacts of uncertainty on trading activity. Stocks that are trading the least are the most influenced by periods of uncertainty in comparison to stocks that are trading at much higher volumes.

This contradicts the findings by Brown (1999) and Antweiler and Frank (2005) who find that previous day's volatility induces an increase in trading volume because volatility signals a higher level of disagreement. A higher level of disagreement implies that more investors are buying and selling without an agreed consensus which should increase trading volume. As this analysis does not include a measure of agreement or disagreement it is not possible to make inference of volatility based on collective behaviour.

As the direction of the coefficient suggest, the risk averse investor is likely to avoid trading at all during periods of uncertainty. Although this effect is pronounced for lower trading stocks it could be indicative of investors being more risk averse to firms which are less established and do not have as much market presence compared to larger and well-known stocks. The stratified results will make this clearer by breaking down effects based on firm size.

TOP 25% SUB-SAMPLE

We turn our attention to the top 25% sub-sample; this includes those with the highest market capitalisations in the overall sample of stocks. On average the impact of lagged sentiment has decreased in comparison to the full sample. This is in line with expectations as firms occupying this sub-sample are unlikely to be

swayed by discussion on Reddit. This is because there are many outside influences that are not controlled for which would take precedence in determining their trading volume. Additionally, there is not much meaningful variation in coefficient sizes across the quantiles. These results are not surprising given the size of the firms in question.

The impact of the change in submissions volume tells a different story, with a wide range of impact on either side of the OLS. A 1% increase in the growth rate of submissions increases trading volume between range of 0.039% to 0.128%, which shows significant heterogeneity.

MIDDLE 50% SUB-SAMPLE

The next sub-sample includes those above the 25% but below 75% in terms of the size of market capitalisation within the sample. Here the direction and significance of the coefficients for both previous week sentiment and growth in submission volumes are similar to the top 25% sample. Although the distinction is seen in the magnitude of the coefficients, at the OLS level and across quantiles coefficients are large. This is in line with the fact that these firms may have fewer external influences by comparison. Therefore a larger portion of their trading volume stems from discussion on WallStreetBets giving rise to larger effects.

Volatility also has a larger impact and bears more significance compared to the prior sample with a very pronounced heterogeneity in coefficients across the quantiles. Drawing a comparison between this and the top 25% sample supports the earlier claim that investors may be more risk averse to firms with a smaller market cap.

BOTTOM 25% SUB-SAMPLE

The final sub-sample reports larger coefficients over the range of quantiles for prior week sentiment, which again are larger at the top of the distribution. A somewhat surprising finding is the lack of significance across the entire range of coefficients including the OLS. This may be plausible due to sample selection. Firms occupying this sub-sample are smaller and not as popular and therefore there are fewer threads of discussion from which to draw sentiment, as such there may not be enough data to recover a meaningful result.

The growth in submissions volume follows a significant and similar pattern to the previous sub-samples ranging from 0.15% to 0.31% in impact for each 1% increase in the growth rate of submissions.

These coefficients are the largest amongst all the sub-samples. Firms in this sub-sample are less valuable and smaller, and less likely to be discussed compared to those in the previous sub-samples. In the event that WallStreetBets discussion leads to further investment into the stock, the marginal impact on trading volume would naturally be higher. This is because there are fewer sources of external variation which influence buying and sell-

ing decisions compared to a larger firm.

Supporting the earlier claim, volatility at the average level has the largest coefficient compared to the sub-samples of higher valued firms. This could be in line with the idea that firms in this sub-sample bear the most risk for an investor and times of uncertainty leads investors to abstain from buying/selling stocks.

DISCUSSION

At a glance, the findings show that WallStreetBets activity (both volume and sentiment) is more pronounced for smaller firms which falls in line with the expectations and justified by the previously mentioned remarks. On the other hand, a common pattern emerges across all the analyses, which is the impact of WallStreetBets behaviour being an increasing effect across quantiles. This is true for both measures of collective behaviour significant or not, which somewhat supports our hypothesis. This effectively claims that the stocks which are experiencing the most volume of trade regardless of their size are the ones which are most influenced by text sentiment and growth in submissions. This is consistent across each band of firm size, which could imply that firms of all sizes are being influenced by WallStreetBets discussion, for both volume and underlying sentiment.

A possible reason for this pattern could be due to the following. A theory presented by Sabherwal et al. (2011) is the pump and dump scheme, often beginning with a few individuals who drive excitement about a stock with very little fundamental underpin, this relies on herd mentality to initiate and spread discussion about this stock. With enough participation, this becomes a self-fulfilling prophecy as myopic investors flock to buy the stock, driving the price up which generates significant returns for the initial investors, who then offload their holdings to reap profits when the stock reaches a critical mass, Bommel (2003).

In the context of this study's findings, the increasing impact of WallStreetBets behaviour across quantiles may be characteristic of a pump and dump operation. This is because high-volume stocks see more variation due to discussion, which may be evidence of the herding mechanism. As volume begins to increase so does discussion because they both act as signals of market participation. The only qualm about this claim is that although pump and dump activities are generally observed amongst smaller companies with weak fundamentals, the described pattern is seen in all three sub-samples which include some of the largest firms in the world. This does not disregard the possibility of it being a pump and dump scheme, however additional experiments using more involved models of social contagion in WallStreetBets discussion would be more appropriate to justify this.

A more plausible explanation could be that stocks towards the bottom of the distribution are unlikely to exhibit much

meaningful variation in volume on average. If there are spikes of higher volume, the stark difference is likely to be influenced by a change in a measure of collective behaviour. Therefore a significant impact of either variable would see more impact on the periods of increased trading volume rather than the lower periods.

6 ROBUSTNESS CHECKS

To ensure the results are robust to changes in sample and specification, we have added three additional checks to describe the relationship between WallStreetBets behaviour and trading volumes. The first check restrict the sample by using a higher cut-off of 200 for the number of submissions. The second check involves restricting the data from the January 2021 short-squeeze event. This is because there is a possibility that the information for firms involved in the short squeeze may contaminate the results and provide an overestimate of the impacts that collective discussion has had on trading volumes. Therefore, as an evaluation of robustness we run additional regressions to exclude these firms from the sample to control for the absence of abnormal events that have taken place. The final check includes an interaction term between the two measures of user sentiment for added interpretation.

STOCKS WITH GREATER THAN 200 SUBMISSIONS

The first result involves a sample reduction, initially the cut-off for the number of submissions for a given stock was 100 for the original analysis. By increasing this cut-off to 200 submissions, the number of observations decrease. Those remaining in the sample are firms which have had more discussion over the year. It is possible that the stocks with less discussion may add nothing more than noise if there is not enough data to extract value from.

The results in table 7 maintain a positive and statistically significant relationship for lagged sentiment and growth in submissions. In comparison to table 3, the coefficient sizes have decreased for lagged sentiment but have increased for growth in submissions, although these changes are marginal.

EXCLUDING OUTLIERS

The second experiment excludes a sample of ‘GME’, ‘AMC’, ‘BB’, ‘NOK’ and ‘KOSS’ stocks. These stocks were subject to a significant quantity of trading in January 2021 during the short squeeze event and subsequently given a temporary halt by trading platforms. Excluding them can allow evaluation of the sentiment relationship under normal market conditions, as they are likely to exhibit some exceptional movement that could skew the results.

Table 8 presents these results; the general pattern remains the same with statistically significant and positive coeffi-

cients across quantiles. Coefficients for lagged sentiment remain similar across quantiles in magnitude and direction, although there is a loss of significance. The growth in submissions volume coefficients have a more noticeable decrease, however the direction of effect and overall impact are largely unaffected. Implying that even under normal circumstances, discussion may be influential in the estimation of trade volumes.

ADDITION OF INTERACTION TERM

The final robustness check presented in table 9 extends the specification by introducing an interaction term between the two measures of collective behaviour. Table 9 presents the results where $S_{it-1} \cdot \Delta A_{it}$ is the variable of interest. This interaction term has a negative direction of effect across all quantiles including the OLS. The implication being that a positive change in both variables will reduce their effective impact on trading volumes.

The impact can be explained using a simple thought experiment, assuming a positive change in submissions growth for both scenarios. For the first scenario, given that there is a positive value for lagged sentiment, there will be an increase in the trading volume, however the negative effect of the interaction will reduce this overall increase. The second scenario has a negative value for lagged sentiment, which will lead to a positive movement in the overall trading volume. This demonstrates the asymmetric behaviour that investors portray in their buying and selling decisions. Both scenarios assume that there is growth in discussion, positive sentiment may be representative of investors buying but it also captures the effect of investors holding onto their assets, which in this case is to do nothing. On the other hand, negative sentiment would likely induce selling which increases the overall trading volume as investors aim to cut their losses. For the risk averse investor, the decision to buy may involve some hesitation. Selling in the face of potential downside is usually a haste decision which results in a lower and higher trading volume for both scenarios respectively. This is built upon the idea of prospect theory, where losses provide more disutility to an investor than utility obtained from gains of an equal amount, Kahneman and Tversky (1979). This could be informative in describing the findings in this extended specification.

7 CONCLUSION

Quantile regression approaches to modelling online user sentiments have been rather scarce, especially using Reddit as a source of collective behaviour. This study attempts to model the heterogeneous relationship between the growth in discussion and the underlying sentiment of the discussion upon the trading volumes of WallStreetBets's commonly discussed stocks from NASDAQ and the NYSE. The panel is modelled using a Method of Moments

Quantile Regression approach which ventures beyond the mean in comparison to Semenova and Winkler (2021) who conduct a similar study. To the best of our knowledge, this is the only analysis which has explored the heterogeneity in social media sentiments using quantile regression, especially using WallStreetBets as a source of collective behaviour.

Results show a positive and significant relationship between collective behaviour and trading volume. The most potent variation occurs amongst stocks in the lowest sub-sample of market capitalisation. The heterogenous effect shows an increasing impact across quantiles of trading volumes for both variables of interest across all sub-samples. Results are robust against restrictions in sample sizes, this provide confidence in the findings. A change in specification reveals a further relationship between the measures of sentiment providing insight into the dynamics between the the measures of collective behaviour.

Overall, this paper has provided interesting contributions to the existing literature. By exploring the heterogenous effect, where a key finding is the larger variation in trading volumes for firms with lower valuations. It could be worthwhile to implement metrics pertaining to Reddit behaviour into existing forecasting models, especially the growth in submission volumes, as this had the most distinct impact.

Financial data is generally observed in a high frequency environment, daily and intraday intervals are not uncommon and therefore the use of weekly data is likely to be a limiting factor for recovering precise estimates of sentiment impacts. This is accentuated by the fact that conversation spreads quickly and there is likely to be more meaningful variation in market behaviour captured through daily data. However in this study the weekly data absorbs this information into a single average. Using inter or even intra-day data would be a technically demanding task as it would require further rules and assumptions to be established surrounding the treatment of out of trading hours conversation.

Another important consideration is regarding the content of the Reddit data. In this study we have limited the data to the submissions made by users, which is of lower volume compared to comments. For each submission, there will be rife conversation by other users which can capture a richer form of sentiment not included in the submission itself. So despite being a more meaningful source of sentiment, it is far more computationally intensive to obtain considering the number of stocks and the length of time being observed. Access to conversation level data would be useful for future work which involves additional variables developed in previous literature, such as an agreement index.

No matter how frequent or precise the data is, inference will ultimately depend on the quality of the sentiments that are

drawn from the analyzer. In this analysis, we have used the default lexicon that VADER provides, even though modification is straightforward, any additional adjustments require human validation using the procedure outlined in, Hutto and Gilbert (2014).

Considering the tone of language frequently used on WallStreetBets, VADER could encounter some limitations to the extent that it may not recognise commonly used jargon. To exemplify this, the terms "paper hands" and "diamond hands" are often used to describe an individual who is holding their assets in the face of losses or an individual who sells at the first sign of losses respectively. These terms signal intent to other users and may play an important role in their subsequent buying or selling behaviours.

These terms are not explicitly included in the lexicon, therefore any instances of these phrases would be ignored, which may skew the classification of the true opinion contained in the message. Further work may benefit by updating the lexicon to include domain specific jargon, all things considered VADER remains a competitive and efficient tool in extracting sentiments.

Finally, an overlooked element of the nature of WallStreetBets discussion is the original source of discussion, Reddit or any online forum for that matter is not an isolated community. Information gathered from another source may escalate discussion on WallStreetBets, however this does not imply that all variation for a given stock's trading volume has been influenced by WallStreetBets and WallStreetBets alone. For example, a Twitter discussion about a stock may percolate into Reddit and subsequently influence the trading volume of the stock. Based on this specification, there could be a potential to overestimate the impact that Reddit can have. In actuality a lot of trading activity had also been derived from Twitter. As a result, it may be valuable to control for the variation caused by discussion in other communities, which may include other subreddits or even other social media. Whilst a demanding task, there is potential to uncover more precise estimates.

BIBLIOGRAPHY

Antweiler, W., & Frank, M. (2005). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259,1294. doi: 10.1111/j.1540-6261.2004.00662.x

Baker, M., & Stein, J. C. (2004). Market liquidity as a sentiment indicator. *Journal of Financial Markets*, *7*(3), 271,299. doi: 10.1016/j.finmar.2003.11.005

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645–1680. doi: 10.1111/j.1540-6261.2006.00885.x

Black, F. (1986). Noise. *The Journal Of Finance*, *41*(3), 528,543. doi: 10.1111/j.1540-6261.1986.tb04513.x

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1,8. doi: 10.1016/j.jocs.2010.12.007

Bommel, J. V. (2003). Rumors. *The Journal of Finance*, *58*(4), 1499-1520. doi: 10.1111/1540-6261.00575

Brown, G. W. (1999). Volatility, sentiment, and noise traders. *Financial Analysts Journal*, *55*(2), 82,90. Retrieved from `https://www.jstor.org/stable/4480157`

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance*, *66*(5), 1461,1499. doi: 10.1111/j.1540-6261.2011.01679.x

Deng, S., Huang, Z., Sinha, A., & Zhao, H. (2018). The interaction between microblog sentiment and stock return: An empirical examination. *MIS Quarterly*, *42*(3), 895,918. Retrieved from `https://ssrn.com/abstract=3054906`

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383,417. doi: 10.2307/2325486

Ganesh, A., & Iyer, S. (2021). Impact of firm-initiated tweets on stock return and trading volume. *Journal of Behavioral Finance*. doi: 10.1080/15427560.2021.1949717

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In *Proceedings of the eleventh acm sigkdd international conference on knowledge discovery in data mining* (pp. 78,87). doi: 10.1145/1081870.1081883

Harris, M., & Raviv, A. (1993). Differences of opinion make a horse race. *The Review of Financial Studies*, *6*(3), 473,506. doi: 10.1093/rfs/5.3.473

Huang, D., Jiang, F., Tu, J., & Zhuou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies*, *28*(3), 791,837. doi: 10.1093/rfs/hhu080

Hutto, C., & Gilbert, E. (2014, June). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social*

media (icwsm-14) (Vol. 8, pp. 216,225). Ann Arbor, MI.

Im, K. S., Pesaran, M., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, *115*(1), 53,74. doi: 10.1016/S0304-4076(03)00092-7

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, *27*(4), 1116,1127. doi: 10.1016/j.ijforecast.2010.11.001

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263,292. doi: 10.2307/1914185

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, *91*(1), 74,89. doi: 10.1016/j.jmva.2004.05.006

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33,50. doi: 10.2307/1913643

Kumar, A. (2006). Determinants of stock trading volume: Evidence from indian stock markets. *Indira Gandhi Institute of Development Research (IGIDR) - Economics*. doi: 10.2139/ssrn.947429

Li, T., Dalen, J. V., & van Rees, P. J. (2018). More than just noise? examining the information content of stock microblogs on financial markets. *Journal of Information Technology*, *33*(1), 50,69. doi: 10.1057/s41265-016-0034-2

Lo, A., & Wang, J. (2000, March). *Trading volume: Definitions, data analysis, and implications of portfolio theory* (Working Paper No. 7625). National Bureau of Economic Research. doi: 10.3386/w7625

Ma, C., & Xiao, S. (2018). Investor sentiment and the prediction of stock returns: a quantile regression approach. *Applied Economics*, *50*(50), 5401,5415. doi: 10.1080/00036846.2018.1486993

Machado, J., & Santos-Silva, J. (2018). Xtqreg: Stata module to compute quantile regression with fixed effects. *Statistical Software Components*. Retrieved from `https://ideas.repec.org/c/boc/bocode/s458523.html`

Martinez-Camara, E., m. Teresa Martin-Baldivia, l. Alfonso Urena-Lopez, & Montejo-Raez, A. (2012). Sentiment analysis in twitter. *Natural Language Engineering*, 1,28. doi: 10.1017/S1351324912000332

Michalak, J. (2020). Does pre-processing affect the correlation indicator between twitter message volume and stockmarket trading volume? *Ekonomia I Prawo. Economics and Law*, *19*(4), 739,755. doi: 10.12775/EiP.2020.048

Ni, Z.-X., Wang, D.-Z., & Xue, W.-J. (2015). Investor sentiment and its nonlinear effect on stock returns: New evidence from the chinese stock market based on panel quantile regression model. *Economic Modelling*, *50*, 266,274. doi: 10.1016/j.econmod.2015.07.007

Pennebaker, J., Booth, R., Boyd, R., & Francis, M. (2015). Linguistic inquiry

and word count [Computer software manual]. Austin, TX. Retrieved from https://downloads.liwc.net.s3.amazonaws.com/LIWC2015_OperatorManual.pdf

Philip, S., Robert, B., Zvi, N., & Daniel, O. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484. doi: 10.1145/1461551.1461583

Sabherwal, S., Sarkar, S. K., & Zhang, Y. (2011). Do internet stock message boards influence trading? evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance and Accounting*, *38*(9-10), 1209,1237. doi: 10.1111/j.1468-5957.2011.02258.x

Semenova, V., & Winkler, J. (2021). Reddit's self-organised bull runs. In *Munich personal repec archive.*

Sprenger, T. O., Tumasjan, A., g. Sandner, P., & Welpe, I. M. (2013). Tweets and trades: the information content of stock microblogs. *European Financial Management*, *20*(5), 926,957. doi: 10.1111/j.1468-036X.2013.12007.x

Swamy, V., & Dharani, M. (2020). Google search intensity and the investor attention effect: A quantile regression approach. *Journal of Quantitative Economics*, *18*(2), 402,423. doi: 10.1007/s40953-019-00185-9

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139,1168. doi: 10.1111/j.1540-6261.2007.01232.x

Vlastakis, N., & N.Markellos, R. (2012). Information demand and stock market volatility. *Journal of Banking Finance*, *36*(6), 1808,1821. doi: 10.1016/j.jbankfin.2012.02.007

Wang, G., Yu, G., & Shen, X. (2021). The effect of online investor sentiment on stock movements: An lstm approach. *Complexity*, *2020*. doi: 10.1155/2020/4754025

APPENDIX

A.1: VARIABLES

INDICES

$i \in [1, 110],\ d \in [1, 252],\ t \in [1, 52]$

Weekly price, returns, volatility and autocovariance from daily data :

p_{id} = Closing Price,

$P_{it} = E\Big[\sum_{t=1}^{52} p_{id} | d \in t\Big]$,

$R_{it} = E\Big(\frac{P_{it}}{P_{it-1}}\Big)$,

$\sigma_{it} = E\big(\sqrt{Var(r_{id}|d \in t)}\ \big),$ where $r_{id} = E\Big(\frac{p_{id}}{p_{id-1}}\Big)$,

$Cov_{it}^{R} = cov(R_{it}, R_{it-1})$.

Weekly dollar trading volume from daily stock volume :

v_{id} = Stock Trading Volume,

$V_{it} = \ln\big(E\big[p_{id} \cdot v_{id} | d \in t\big]\big)$.

Weekly sentiment from daily sentiment values :

s_{id} = Average Daily Sentiment $\in (0, 1)$,

$S_{it} = E[s_{id} | d \in t] \in (0, 1)$,

Weekly submissions volume from daily submissions :

a_{id} = Volume of submissions,

$A_{it} = \ln\big(\sum a_{id} | d \in t\big)$,

PROGRAMMING LANGUAGES & TOTAL FILE SIZE

XTQREG command used for quantile regression in STATA, Machado and Santos-Silva (2018)

Python (.ipynb): 82kb

STATA (.do): 15kb

R (.R): 6kb

A.2: TABLES

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Max
S_{it-1}	5,460	0.266814	0.375984	-0.9981	1
ΔA_{it}	5,460	0.021123	1.006522	-5.11799	5.257495
V_{it-1}	5,460	18.40975	2.409176	8.163006	24.7466
σ_{it-1}	5,460	0.048909	0.06606	0.000875	2.132209
Cov_{it}^{R}	5,570	0.550344	1.748157	-12.4728	28.15126

Table 2: Im-Pesaran-Shin Unit Root Test

Variable	Test-statistic	p-value
S_{it-1}	-48.18	0.00
ΔA_{it}	-79.25	0.00
V_{it-1}	-18.20	0.00
σ_{it-1}	-77.21	0.00
Cov_{it}^{R}	-57.75	0.00

Table 3: Full Sample Results

		Quantile Probability						
Dependent: V_{it}	OLS	0.1	0.2	0.4	0.5	0.6	0.8	0.9
S_{it-1}	0.111***	0.0733*	0.0842**	0.0992***	0.106***	0.114***	0.135***	0.156**
	(0.0246)	(0.0347)	(0.0274)	(0.0222)	(0.0227)	(0.0257)	(0.0400)	(0.0581)
ΔA_{it}	0.180***	0.0960***	0.120***	0.154***	0.170***	0.188***	0.235***	0.283***
	(0.00846)	(0.0134)	(0.0106)	(0.00862)	(0.00885)	(0.00999)	(0.0155)	(0.0225)
V_{it-1}	0.858***	0.947***	0.921***	0.885***	0.868***	0.849***	0.799***	0.749***
	(0.00777)	(0.0216)	(0.0171)	(0.0139)	(0.0142)	(0.0161)	(0.0250)	(0.0363)
σ_{it-1}	-1.401***	-3.074***	-2.589***	-1.914***	-1.595**	-1.229*	-0.293	0.655
	(0.150)	(0.785)	(0.622)	(0.504)	(0.516)	(0.583)	(0.906)	(1.317)
Cov_{it}^{R}	0.0491***	0.0360***	0.0398***	0.0451***	0.0476***	0.0505***	0.0578***	0.0652***
	(0.00552)	(0.00814)	(0.00645)	(0.00522)	(0.00534)	(0.00604)	(0.00939)	(0.0136)
Constant	2.647***							
	(0.139)							
Observations	5460	5460	5460	5460	5460	5460	5460	5460

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Top 25% Market Cap

		Quantile Probability						
Dependent: V_{it}	OLS	0.1	0.2	0.4	0.5	0.6	0.8	0.9
S_{it-1}	0.0858***	0.0735	0.0772*	0.0822***	0.0845***	0.0873**	0.0939*	0.100
	(0.0250)	(0.0396)	(0.0312)	(0.0247)	(0.0247)	(0.0274)	(0.0412)	(0.0588)
ΔA_{it}	0.0800***	0.0392**	0.0516***	0.0680***	0.0758***	0.0850***	0.107***	0.128***
	(0.00909)	(0.0145)	(0.0115)	(0.00916)	(0.00917)	(0.0101)	(0.0152)	(0.0216)
V_{it-1}	0.621***	0.663***	0.650***	0.633***	0.625***	0.616***	0.594***	0.572***
	(0.0217)	(0.0373)	(0.0294)	(0.0234)	(0.0234)	(0.0259)	(0.0390)	(0.0555)
σ_{it-1}	-0.379*	-0.569	-0.512	-0.435	-0.399	-0.356	-0.254	-0.155
	(0.150)	(0.380)	(0.300)	(0.238)	(0.238)	(0.263)	(0.396)	(0.565)
Cov_{it}^{R}	0.0178***	0.0219***	0.0206***	0.0190***	0.0182***	0.0173***	0.0150**	0.0129
	(0.00352)	(0.00467)	(0.00368)	(0.00292)	(0.00292)	(0.00323)	(0.00487)	(0.00693)
Constant	7.840***							
	(0.450)							
Observations	1404	1404	1404	1404	1404	1404	1404	1404

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Mid 50% Market Cap

		Quantile Probability						
Dependent: V_{it}	OLS	0.1	0.2	0.4	0.5	0.6	0.8	0.9
S_{it-1}	0.100**	0.0837	0.0886*	0.0951**	0.0984**	0.102**	0.111*	0.120
	(0.0335)	(0.0557)	(0.0445)	(0.0349)	(0.0342)	(0.0366)	(0.0537)	(0.0790)
ΔA_{it}	0.191***	0.115***	0.137***	0.168***	0.183***	0.199***	0.240***	0.283***
	(0.0116)	(0.0214)	(0.0171)	(0.0135)	(0.0132)	(0.0142)	(0.0208)	(0.0306)
V_{it-1}	0.873***	0.971***	0.942***	0.902***	0.883***	0.862***	0.809***	0.753***
	(0.0101)	(0.0317)	(0.0254)	(0.0200)	(0.0196)	(0.0211)	(0.0309)	(0.0454)
σ_{it-1}	-1.789***	-3.800**	-3.205**	-2.396**	-1.993*	-1.580	-0.500	0.650
	(0.220)	(1.367)	(1.093)	(0.859)	(0.841)	(0.902)	(1.322)	(1.947)
Cov_{it}^R	0.112***	0.0911***	0.0972***	0.106***	0.110***	0.114***	0.125***	0.137***
	(0.0111)	(0.0223)	(0.0178)	(0.0140)	(0.0137)	(0.0147)	(0.0215)	(0.0317)
Constant	2.344***							
	(0.177)							
Observations	2730	2730	2730	2730	2730	2730	2730	2730

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Bottom 25% Market Cap

		Quantile Probability						
Dependent: V_{it}	OLS	0.1	0.2	0.4	0.5	0.6	0.8	0.9
S_{it-1}	0.135	0.0162	0.0445	0.0981	0.116	0.144*	0.213	0.282
	(0.0709)	(0.0945)	(0.0779)	(0.0623)	(0.0641)	(0.0731)	(0.116)	(0.169)
ΔA_{it}	0.224***	0.150***	0.168***	0.201***	0.212***	0.229***	0.272***	0.314***
	(0.0227)	(0.0336)	(0.0277)	(0.0223)	(0.0229)	(0.0261)	(0.0412)	(0.0602)
V_{it-1}	0.846***	0.909***	0.894***	0.866***	0.856***	0.842***	0.806***	0.769***
	(0.0169)	(0.0314)	(0.0259)	(0.0208)	(0.0214)	(0.0244)	(0.0386)	(0.0563)
σ_{it-1}	-2.320***	-3.672***	-3.351***	-2.744***	-2.535***	-2.227**	-1.445	-0.664
	(0.378)	(0.892)	(0.736)	(0.590)	(0.606)	(0.691)	(1.093)	(1.595)
Cov_{it}^R	0.606***	0.533***	0.550***	0.583***	0.594***	0.611***	0.653***	0.695**
	(0.0515)	(0.129)	(0.106)	(0.0851)	(0.0875)	(0.0998)	(0.158)	(0.230)
Constant	2.592***							
	(0.267)							
Observations	1326	1326	1326	1326	1326	1326	1326	1326

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Robustness Check: Greater Than 200 Submissions

		Quantile Probability						
Dependent: V_{it}	OLS	0.1	0.2	0.4	0.5	0.6	0.8	0.9
S_{it-1}	0.119***	0.0793	0.0912**	0.107***	0.115***	0.124***	0.145**	0.167*
	(0.0309)	(0.0445)	(0.0345)	(0.0285)	(0.0303)	(0.0354)	(0.0545)	(0.0786)
ΔA_{it}	0.168***	0.0903***	0.114***	0.145***	0.161***	0.178***	0.218***	0.261***
	(0.00990)	(0.0160)	(0.0124)	(0.0103)	(0.0109)	(0.0128)	(0.0196)	(0.0283)
V_{it-1}	0.868***	0.955***	0.929***	0.894***	0.876***	0.856***	0.811***	0.764***
	(0.00940)	(0.0264)	(0.0205)	(0.0170)	(0.0181)	(0.0211)	(0.0324)	(0.0467)
σ_{it-1}	-1.211***	-2.612**	-2.194***	-1.630**	-1.342*	-1.033	-0.310	0.454
	(0.163)	(0.802)	(0.623)	(0.514)	(0.547)	(0.640)	(0.984)	(1.418)
Cov_{it}^R	0.0412***	0.0305***	0.0337***	0.0380***	0.0402***	0.0426***	0.0482***	0.0540***
	(0.00573)	(0.00806)	(0.00625)	(0.00516)	(0.00548)	(0.00642)	(0.00987)	(0.0142)
Constant	2.551***							
	(0.176)							
Observations	3339	3339	3339	3339	3339	3339	3339	3339

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Robustness Check: Exclusion of Extremities

		Quantile Probability						
Dependent: V_{it}	OLS	0.1	0.2	0.4	0.5	0.6	0.8	0.9
S_{it-1}	0.108***	0.0763*	0.0856**	0.0985***	0.105***	0.112***	0.130**	0.148*
	(0.0250)	(0.0354)	(0.0280)	(0.0226)	(0.0231)	(0.0260)	(0.0402)	(0.0586)
ΔA_{it}	0.174***	0.0932***	0.117***	0.149***	0.165***	0.183***	0.228***	0.274***
	(0.00867)	(0.0137)	(0.0109)	(0.00880)	(0.00900)	(0.0101)	(0.0157)	(0.0228)
V_{it-1}	0.853***	0.948***	0.920***	0.882***	0.863***	0.843***	0.789***	0.735***
	(0.00795)	(0.0226)	(0.0180)	(0.0146)	(0.0149)	(0.0167)	(0.0259)	(0.0377)
σ_{it-1}	-1.330***	-3.016***	-2.527***	-1.847***	-1.517**	-1.156	-0.211	0.747
	(0.154)	(0.841)	(0.667)	(0.538)	(0.550)	(0.618)	(0.958)	(1.394)
Cov_{it}^R	0.0463***	0.0334***	0.0371***	0.0423***	0.0448***	0.0476***	0.0548***	0.0621***
	(0.00573)	(0.00787)	(0.00624)	(0.00503)	(0.00514)	(0.00578)	(0.00895)	(0.0130)
Constant	2.739***							
	(0.142)							
Observations	5256	5256	5256	5256	5256	5256	5256	5256

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Robustness Check: Inclusion of Interaction Term

		Quantile Probability						
Dependent: V_{it}	OLS	0.1	0.2	0.4	0.5	0.6	0.8	0.9
$S_{it-1} \cdot \Delta A_{it}$	-0.0682**	-0.0528	-0.0573	-0.0635**	-0.0664**	-0.0697*	-0.0784	-0.0871
	(0.0242)	(0.0379)	(0.0298)	(0.0242)	(0.0250)	(0.0283)	(0.0446)	(0.0649)
S_{it-1}	0.125***	0.0962**	0.105***	0.116***	0.121***	0.127***	0.143***	0.160**
	(0.0256)	(0.0349)	(0.0274)	(0.0223)	(0.0230)	(0.0261)	(0.0411)	(0.0598)
ΔA_{it}	0.197***	0.111***	0.136***	0.170***	0.187***	0.205***	0.253***	0.302***
	(0.0118)	(0.0195)	(0.0154)	(0.0125)	(0.0129)	(0.0147)	(0.0230)	(0.0336)
V_{it-1}	0.852***	0.947***	0.920***	0.881***	0.863***	0.843***	0.789***	0.736***
	(0.00794)	(0.0227)	(0.0179)	(0.0146)	(0.0151)	(0.0171)	(0.0269)	(0.0391)
σ_{it-1}	-1.309***	-3.011***	-2.515***	-1.833***	-1.506**	-1.146	-0.188	0.775
	(0.154)	(0.844)	(0.664)	(0.540)	(0.558)	(0.634)	(0.996)	(1.450)
Cov_{it}^R	0.0461***	0.0329***	0.0368***	0.0421***	0.0446***	0.0474***	0.0548***	0.0623***
	(0.00573)	(0.00786)	(0.00619)	(0.00502)	(0.00519)	(0.00589)	(0.00927)	(0.0135)
Constant	2.740***							
	(0.142)							
Observations	5256	5256	5256	5256	5256	5256	5256	5256

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: List of Firms in Sample

Ticker Symbol	**Company Name**
AAL	American Airlines Group Inc
AAPL	Apple Inc
ACB	Aurora Cannabis Inc
ADMP	Adamis Pharmaceuticals Corporation
AG	First Majestic Silver Corp
ALLY	Ally Financial Inc
AMC	AMC Entertainment Holdings Inc
AMCX	AMC Networks Inc
AMD	Advanced Micro Devices Inc
AMZN	Amazon.com Inc
ATH	Athene Holding Ltd
ATOS	Atossa Therapeutics Inc
BABA	Alibaba Group Holding Limited
BB	BlackBerry Limited
BBBY	Bed Bath & Beyond Inc
BBW	Build-A-Bear Workshop Inc
BBY	Best Buy Co. Inc
BCRX	BioCryst Pharmaceuticals Inc
BIO	Bio-Rad Laboratories Inc
BNGO	Bionano Genomics Inc
CAT	Caterpillar Inc
CCL	Carnival Corporation & plc
CGC	Canopy Growth Corporation
CIDM	Cinedigm Corp
CLOV	Clover Health Investments Corp
CLVS	Clovis Oncology Inc
CRSR	Corsair Gaming Inc
CTRM	Castor Maritime Inc
DIS	The Walt Disney Company
DKNG	DraftKings Inc
DM	Desktop Metal Inc
ETSY	Etsy Inc
EXP	Eagle Materials Inc
FB	Facebook Inc
FCEL	FuelCell Energy Inc
FIZZ	National Beverage Corp
FUBO	fuboTV Inc
GE	General Electric Company
GEVO	Gevo Inc
GM	General Motors Company
GME	GameStop Corp
GNUS	Genius Brands International Inc
GOEV	Canoo Inc
GOGO	Gogo Inc
GPRO	GoPro Inc
GSAT	Globalstar Inc
GT	The Goodyear Tire & Rubber Company
GTE	Gran Tierra Energy Inc
HOOD	Robinhood Markets Inc. Class A Common Stock
HYLN	Hyliion Holdings Corp
IBIO	iBio Inc
IBKR	Interactive Brokers Group Inc
IDEX	Ideanomics Inc
IVR	Invesco Mortgage Capital Inc
JAGX	Jaguar Health Inc
JPM	JPMorgan Chase & Co
KODK	Eastman Kodak Company
KOSS	Koss Corporation
LGND	Ligand Pharmaceuticals Incorporated
MA	Mastercard Incorporated
MARA	Marathon Digital Holdings Inc
MSFT	Microsoft Corporation
MT	ArcelorMittal
MVIS	MicroVision Inc
NAK	Northern Dynasty Minerals Ltd
NIO	NIO Inc

NKLA	Nikola Corporation
NNDM	Nano Dimension Ltd
NOK	Nokia Corporation
NVAX	Novavax Inc
OCGN	Ocugen Inc
OGI	OrganiGram Holdings Inc
OPK	OPKO Health Inc
PFE	Pfizer Inc
PLTR	Palantir Technologies Inc
PLUG	Plug Power Inc
PRTY	Party City Holdco Inc
QS	QuantumScape Corporation
RIG	Transocean Ltd
RIOT	Riot Blockchain Inc
RKT	Rocket Companies Inc
SENS	Senseonics Holdings Inc
SIRI	Sirius XM Holdings Inc
SIX	Six Flags Entertainment Corporation
SKIN	The Beauty Health Company
SKT	Tanger Factory Outlet Centers Inc
SNAP	Snap Inc
SNDL	Sundial Growers Inc
SNOW	Snowflake Inc.
SOFI	SoFi Technologies Inc. Common Stock
SPCE	Virgin Galactic Holdings Inc
SPWR	SunPower Corporation
SRNE	Sorrento Therapeutics Inc
TD	The Toronto-Dominion Bank
TDA	Telephone & Data Systems Inc
TLRY	Tilray Inc
TNXP	Tonix Pharmaceuticals Holding Corp
TR	Tootsie Roll Industries Inc
TSLA	Tesla Inc
TXMD	TherapeuticsMD Inc
UBER	Uber Technologies Inc
UWMC	UWM Holdings Corp
VXRT	Vaxart Inc
WISH	ContextLogic Inc
WKHS	Workhorse Group Inc
X	United States Steel Corporation
XL	XL Fleet Corp
XSPA	XpresSpa Group Inc
Z	Zillow Group Inc
ZOM	Zomedica Corp