CAN SOCIAL MEDIA FACILITATE INFORMED TRADING?

A Dissertation

by

YUAN XUE

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Hwagyun Kim
Committee members,	Alexander L. Brown
	Yong Chen
	Wei Wu
Head of Department,	Christa H.S. Bouwman

May 2022

Major Subject: Finance

ABSTRACT

In this paper, we collect data from the online investment platform Seeking Alpha to study the effect of social media on stock market trading. First, we find that the tone of Seeking Alpha articles and comments is informative in predicting future retail orders and short sales. Furthermore, we find that after Seeking Alpha article publications, retail order flows are significantly less contrarian, suggesting that social media accelerates information adoption by retail investors to make less uninformed decisions. Second, we decompose the tone of articles and comments into parts based on the informative signals and noises to find that more positive article noises predict more net buys from retail investors but more short sales from sophisticated investors. Related, predictability of both retail order imbalances and short sales on future returns are significantly greater on days that have Seeking Alpha articles published. Zero-cost portfolios combining information of Seeking Alpha article publications and retail order imbalances (short sales) realize an annualized alpha of around 16% (38%). Overall, our findings suggest that social media facilitate informed trading and help improve market efficiency.

DEDICATION

To my father, in loving memory.

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Kim, and my committee members, Dr. Brown, Dr. Chen, Dr. Wu, for their guidance and support throughout the course of this research.

I would like to thank my elder sister for taking care of my parents during my PhD study.

I would like to thank Boone Bowles, Xiao Cen, Bouwman Christa, Tristan Fitzgerald, Dora Horstman, Shane Johnson, Adam Kolasinski, Marco Rossi, and Nan Yang for their helpful comments and suggestions.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Hwagyun Kim (Committee Chair), Dr. Yong Chen and Dr. Wei Wu of the Department of Finance and Dr. Alexander L. Brown of the Department of Economics.

All work for the dissertation was completed by the student, under the advisement of Hwagyun Kim of the Department of Finance.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

CONTRIBUTORS AND FUNDING SOURCES v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES ix

1 INTRODUCTION 1

2 LITERATURE REVIEW 8

2.1 Using Social Media Data to Predict Stock Market Movement 8
2.2 Retail Investors, Uninformed or Informed? 9
2.3 Retail Investors as Liquidity Providers 10
2.4 Informativeness of Short Sales 11
2.5 Other Researches Using the SA Data Set 12

3 DATA AND DESCRIPTIVE STATISTICS 13

3.1 Seeking Alpha 13
3.2 TAQ Retail Order Flows 15
3.3 CBOE Short Sales 15
3.4 Informational Variables 16
3.4.1 Earnings Level and Earnings Growth 16
3.4.2 IBES Analyst Forecast 16
3.4.3 IBES Recommendations 17
3.4.4 RavenPack News 17
3.5 Returns and Abnormal Returns 17
3.6 Descriptive Statistics 18

4 RETAIL INVESTORS 20

4.1 Social Media Tone and Order Imbalances of Retail Investors 20
4.2 Social Media Talks That Most Predict Retail Order Imbalances 21
4.3 Impact of Social Media on Liquidity Provision of Retail Investors 21
4.3.1 Investor Attention 23

4.4 Impact of Social Media on Return Predictability of Retail Orders 23
4.5 Portfolio Analysis 25
4.5.1 Alphas and Risk Loadings 26

5 SHORT SELLERS 28

5.1 Social Media Tone and Short Sale Flows 28
5.2 Social Media Noises and Short Sale Flows 29
5.3 Social Media Talks That Most Predict Short Sale Levels 30
5.4 Impact of Social Media on Return Predictability of Short Sales 31
5.5 Portfolio Analysis 31
5.5.1 Alphas and Risk Loadings 33

6 INFORMATIONAL EFFICIENCY 34

7 CONCLUSION 37

REFERENCES 38

APPENDIX A FIGURES 41

APPENDIX B TABLES 45

APPENDIX C SUPPLEMENTARY TABLES 58

LIST OF FIGURES

FIGURE Page

A.1 Words (Bigrams) Most Relevant to Retail Order Imbalances 41

A.2 Words (Bigrams) Most Relevant to Short Sale Flows 42

A.3 Evolution of the Informed and Uninformed Retail Portfolios 43

A.4 Evolution of the Informed and Uninformed “Short Seller” Portfolios 44

LIST OF TABLES

TABLE Page

B.1 Descriptive Statistics 45

B.2 Social Media Tone and Retail Order Imbalances 47

B.3 The Impact of Social Media on Liquidity Provision of Retail Investors 48

B.4 The Impact of Social Media Coverage on Return Predictability of Retail Order Imbalances 49

B.5 Alpha and Risk Loading of Two Portfolios Mimicking Retail Investments 50

B.6 Social Media Tone and Short Sale Flows 51

B.7 Projecting Social Media Tone on Informational Variables 52

B.8 Social Media Noises and Short Sale Flows 54

B.9 Impact of Social Media on Return Predictability of Short Sales 55

B.10 Alphas and Risk Loadings of Two portfolios Based on Short Sale Flow 56

B.11 The Impact of Social Media on Informational Efficiency of Stock Prices 57

C.1 List of Variable Definitions 58

C.2 Social Media Noises and Retail Order Imbalances 61

C.3 Return Predictability of Social Media Tone 62

C.4 The Impact of Social Media on Informational Efficiency of Stocks (Binary Variables on RHS) 63

C.5 The Impact of Social Media on Stock Price Volatility 64

1 INTRODUCTION

Social media is one of the most voluminous and easy to access data sources in modern era. Many researchers in both financial industry and academics try to use social media data to construct more efficient models for predicting stock prices and to develop more profitable trading strategies. However, extant research shows weak or mixed evidence on predictability of social media data on future stock price movements. The reason of the weak predictability comes from two aspects. First, stock price movements largely follow the efficient market hypothesis (EMH), under which stock price fully, accurately, and instantly incorporates all available information. Thus information conveyed by social media immediately becomes stale once it is released and loses its predictability on future prices. Second, even if we believe that stock prices are slow at incorporating new information, which enables short-term predictability, social media information is too noisy to be adopted by investors and thus have little material impact on future prices.

In less efficient market where stock prices slowly adjust on new information, a piece of social media message containing new information of a stock's fundamental value can materialize its implication on subsequent stock prices only when investors on the market believe so and act on the information. Hence, to justify that social media is useful in predicting future stock returns, one should first empirically show that investors do trade on social media information and do benefit from doing so. Moreover, empirical studies on the interactions between investor trading and social media information also possibly provide us detailed knowledge about what types of social media information investors mostly trade on and what types are most profitable to trade on, which in turn helps improve our trading strategies based on social media data.

Therefore, in this paper we study how equity market investors react to social media information through trading, and its asset-pricing implications. Specifically, we study how the two types of investors, the retail investors and short sellers, use social media information and whether this make their tradings more informed. In literature, retail investors are often taken as unsophisticated, uninformed, and even noise traders, whereas short sellers are taken as sophisticated, informed, and rational investors. Consistent with this conventional idea, we find that both types of investors can benefit from social media information, but they use social media data with different levels of sophistication. To retail investors, social media is an important channel to supplement their scarce information sources and enable them to respond in a more timely manner to new information released on equity markets. To short sellers, social media can help them identify investor sentiment in equity markets that is not supported by fundamentals, and implement tradings that are in the opposite direction of retail tradings. Based on the above findings, we constructed portfolios by combining the social media data with the data of order flows of retail investors and short sellers, which realize significantly higher returns than that using only social media data or order flow data alone.

We obtain our social media data by scraping web pages of Seeking Alpha (hereafter SA), which is one of the largest online investment communities. We construct a panel data set suitable for cross-sectional analysis using the stock ticker tags and timestamps of the web pages. The observations of the panel data are firm-day level, and the sample period is from 2007 to 2019. We augment this data set with data from the Trades and Quotes (TAQ) database and the Cboe Global Markets, Inc.(Cboe) for the order flows of retail investors and short sellers respectively.

In the first section of the empirical analyses, we study how social media interconnect with retail trading. As a preliminary piece of evidence that social media information is relevant to retail trading, we first show that the tone of social media is informative in predicting retail order flows. Our panel regression result shows that the negativeness of SA articles is positively associated with net sells by retail investors in contemporaneous and subsequent trading days. Further, to study what contents of social media message retail investors mostly react to, we use Supporting Vector Classifier (SVC) model to extract from the SA articles the words (bigrams) that are most relevant to retail order flows of next trading day. We find that the terms most relevant to future retail net buys, such as `surge`, `synergy`, `success`, `efficiency`, `poise`, `competitive_advantage`, tend to be positive, and terms most relevant to future retail net sells, such as `division`, `restructuring`, `decline`, `bad`, `problem`, tend to be negative.

Through studying the effect of social media on liquidity provision of retail investors, we present a second piece of evidence on the relevance of social media information on retail tradings. The literature has documented (see Section 2.3) that retail investors act as liquidity providers — aggregate retail order imbalances are contrarian. Previous theoretical research shows that it is partly due to retail investors' underreaction to private information contained in the order flows of informed investors. If social media can democratize information access for retail investors, it should primarily move them away from contrarian trading. Indeed, our regression results show that the contrarian trading by retail investors is 30%-50% smaller after the publications of SA articles. We further find that the effect is economically large and statistically significant only when commenters respond to the articles, which indicates that the effect that we find is not merely a news effect, but a result of investors actively obtaining and screening information through social media.

Recent research finds that retail investor order flows predict future returns. This result indicates that retail investors do trade rationally rather than being pure noise traders. We find that information from social media makes retail traders more informed. Regression results show that for the stock-day combinations with no SA articles, one standard deviation increase of net buys by retail investors leads to a 2bps increase of the following day returns, whereas for the stock-day combinations with SA articles, the increase of returns of next trading day is 5bps higher. Our analyses further distinguish retail investors' net buys on SA bullish opinions, and net sells on SA bearish opinions. We show that profits mainly come from net buys on bullish opinions.

Based on our observation that retail investors proactively use and benefit from social media information, we argue that a trading strategy that combines both the retail trading variable and social media variable should perform better than that only uses single variable. We construct two daily-rebalanced portfolios mimicking tradings by retail investors. The first portfolio has access to SA article opinions (informed), whereas the second portfolio does not (uninformed). We compute the monthly evolution of $1 investment on the long and short legs of the two portfolios. The value of the informed, long portfolio grows from $1 to $11.6 in 2018, which drops back to $8.3 at the end of 2018. The value of the corresponding uninformed portfolio grows to $6.6 in 2018 and drops back to $5.3 at the end of 2018. The short legs of both the informed and uninformed portfolios underperform the market portfolio and are close in value. This result shows that the information contained in SA articles helps retail investors to identify undervalued stocks rather than to help them sell overvalued stocks short.

We then regress the daily returns of the two retail portfolios on Fama-French five factors and the extended eight factors. We find that the informed (uninformed) portfolio realizes an annualized alpha of 16% (11%). Moreover, the informed portfolio has a significant loading on market factors, which implies that although opinions of SA articles provide retail investors arbitrage opportunities, they also let the retail investors chase the market trend and hence accumulate systematic risks in their portfolios.

In the second section of the empirical analyses, we study how social media interconnects short selling. Similar to the case of retail investment, we first show that social media is informative in predicting short sales activity. In panel regressions, we find that the negativeness of SA articles is positively associated with contemporaneous and next trading day short sales. Using the SVC model, we build the connections between the contents of social media talks and short sales activity. We find that the relation between the two is not apparent to see. The short sellers seem to be more dedicated to searching for certain types of information from the article, rather than generally following the bullish or bearish opinions of SA articles like the retail investors. Interestingly, we find that the bigram `disclosure_long` is strongly related to later heavy short selling, which implies that short sellers trade against the disclosed long position of SA authors.

Social media can be a channel that provides information of stock fundamentals to investors but can also be a channel that transmits irrational optimism or pessimism (noise) to individual investors about a stock. One interesting question is how arbitrageurs react to irrational sentiment (or noise) on social media? Do they trade against social media sentiment as they believe that stock price will reverse to fundamental shortly? Or do they do nothing because of fear of squeeze of their short positions? Or do they trade in the same direction as the sentiment, for they believe that the present trend will continue? To answer this question, we design a two-step analysis. First, we decompose the opinions of social media into information and noise by regressing the variable of negativeness of SA articles/comments onto a group of informational variables of earnings announcement, analyst revisions, and Ravenpack cash-flow news. We then use the residuals of the regression as a proxy of the noises of social media. Regressing the contemporaneous or next-trading-day short sale variable on the noise variables in the second step, we find that short sellers trade against the noises of SA articles. In contrast, retail order flows are in agreement with both the noises of SA articles and comments.

We then analyze whether and how social media impacts the return predictability of short sales. We find that the predictability is much larger for stock-day combinations that have SA articles published than those do not. This suggests that short sale trading is more informed in days that have SA article publications. Further, we find that short sale trading is more informed only on days when bullish articles are published.

Same as when we analyze retail trading, we find that the trading strategy that combines both variable of trading of short sellers and variable of social media performs better than that using single variable. We construct two zero-cost portfolios using the information of Cboe short sale data. We divide the stocks into long and short legs by the midpoint of adjusted short sales at a daily frequency for both portfolios. For the "informed" portfolio, we require that all the stocks in both long leg or short leg have SA articles published on that day; for the "uninformed" portfolio, we require that none of the stocks have SA articles published. The two portfolios are rebalanced every day. We compute the equal-weighted returns of the two portfolios on the next trading day. We compare the evolution of $1 investment of the long legs and short legs of both portfolios. Starting from $1 at the beginning of 2008, the value of the long leg of the uninformed portfolio grows to $12 at the end of 2019, whereas the long leg of the informed portfolio grows to as high as $58 at the end of 2019. The long position that combines the information of Cboe lightly shorted stocks and SA article publications makes a big profit. Because Cboe posts their short sale data every day after the close of the market, it is also easy to obtain the same day SA article data; we wonder if we can test this strategy in real-world quantitative trading. The short legs of the two portfolios underperform the market portfolio, and their performances are indistinguishable. This finding suggests that social media are more helpful for short-sellers avoiding undervalued stocks than identifying overvalued stocks.

We then regress the daily returns of the two “short-seller” portfolios on Fama-French five factors and the extended eight factors. We find that the informed (uninformed) portfolio realizes an annualized alpha of 38% (14%). Overall, our analysis suggests that short sellers are potentially highly skillful in processing social media information (and noise) for profits.

In the last section of the empirical analyses, we evaluate the impact of social media on the informational efficiency of stock prices. Given our result suggesting that social media can render investors trade informed, we expect it would shorten the duration that stock prices reflect a new bit of information. To test this hypothesis, we use the price-delay measures as dependent variables. We find that price-delays are significantly smaller (larger) for firm-month combinations with more SA article (SA comment) coverage. These results suggest that the SA articles make average investors more informed, but SA comments reflect the market noises. As a robustness check, we also use return volatility as a proxy of informational efficiency, and the result is consistent with the case of the price-delay measure. Moreover, our results are robust after controlling analyst coverage and news volume, which suggests that the contribution of SA articles to market efficiency is incremental.

Taken together, our empirical analyses suggest that social media provides profitable information to both less sophisticated retail investors and more sophisticated short sellers. The bullish articles are most informative, and social media are most useful in identifying the undervalued stocks. However, social media can also transmit noises. Indeed, we observe that retail investors chase the noises, but we also observe that short sellers trade against noises. Trading strategies that combine variables of trading and variable of social media coverage realize significantly positive risk-adjusted returns.

The rest of the paper proceeds as follows. Section 2 reviews the related literature. Section 3 introduces data and provides descriptive statistics. Section 4 analyzes social media's effect on retail investment, and section 5 analyzes social media's effect on short sale activities. Section 6 investigates the impact of social media on informational efficiency. Section 7 concludes.

2 LITERATURE REVIEW

This paper speaks to several strands of literature in social media, retail investors, and short sellers.

2.1 Using Social Media Data to Predict Stock Market Movement

The paper contributes to the literature which explores whether investors' opinions stated on social media are informative of future stock market movement. Antweiler and Frank (2004) use the Naive Bayes algorithm to classify Yahoo! Finance messages to the three categories of buy, hold and sell and aggregate them to a single measure of bullishness of messages. They find that return predictability of bullishness of messages is statistically significant but economically small. They also find that the number of messages significantly predicts market volatility on next day. Das and Chen (2007) construct a system of text processing algorithms to extract sentiment from messages of Yahoo! stock message board. They find that the level of sentiment index of tech-sector stocks weakly predicts the level of stock price index and that message volume significantly explains changes in stock levels as well as volatility. Chen, De, Hu, and Hwang (2014) observe that both the fraction of negative words in SA articles and comments negatively predict the cross section of stock returns. They argue that the return predictability comes from the value-relevant information in SA articles and comments that is not incorporated into stock price yet. They support this argument by showing that SA views prior to earnings announcement predict earnings surprise. Avery, Chevalier, and Zeckhauser (2016) find that the zero-cost portfolio long the positive picks by individual users of MotleyFool and short their negative picks yields 12% annual returns. They find that the profitability of the portfolio mainly comes from the short leg. In this study we construct long-short portfolios that combine the information of SA article posts and the information of order flows of retail investors (or short sales flows). The returns of our portfolios are significantly higher than those using order flow data alone or those using social media data alone documented in prior literature.

2.2 Retail Investors, Uninformed or Informed?

This paper contributes to research about the informativeness of retail tradings. Many studies have concluded that retail investors are overall uninformed. Odean (1999) finds that on average the stocks retail investors buy underperform those they sell. Barber and Odean (2000) find that households that trade frequently earn significantly lower annualized net return than those that trade infrequently. Grinblatt and Keloharju (2001) find that retail investors of Finnish stock market more likely to sell stocks with large positive returns in the recent past and with prices at their monthly highs (disposition effect). Hvidkjaer (2008) finds that stocks favored by retail investors underperform stocks out of favor with retail investors up to two years.McLean, Pontiff, and Reilly (2020b) construct an index based on 130 anomalies and find that retail investors trade against anomalies. However, more recent studies find that retail order imbalances are informative about stock returns over short horizons. Kaniel, Saar, and Titman (2008) document positive excess returns after intense buying by retail investors and negative excess returns after they sell. Kaniel, Liu, Saar, and Titman (2012) find that the abnormal returns of stocks accumulated by retail investors prior earnings announcement exceeds that of stocks sold by retail investors later. Kelley and Tetlock (2013) find that net buying from both market orders and limit orders positively predicts firms' monthly returns with no evidence of return reversal later. Boehmer, Jones, Zhang, and Zhang (2021) develop a new method to identify order flow of retail investors from TAQ data and find that stocks with net retail buys outperform those with net retail sells over the following week. McLean, Pontiff, and Reilly (2020a) find that retail trades are responsive to revisions in analyst recommendations and price targets in the direction of the revision and earn higher returns when they do so. In this paper, we find that retail trades are in agreement with tones of SA articles and comments. Retail order imbalances are more informative in predicting next day stock returns on days with SA articles than days without SA articles. However, we also find evidence that aggregate retail order flows are in agreement with noises of social media.

2.3 Retail Investors as Liquidity Providers

This paper contributes new empirical evidence that helps explain the reason of liquidity provision by retail investors. The phenomenon of liquidity provision by retail investors is well documented in literature (Kaniel, Saar, and Titman (2008), Kelley and Tetlock (2013), Barrot, Kaniel, and Sraer (2016), Boehmer, Jones, Zhang, and Zhang (2021)). Retail order imbalances are contrarian, i.e., aggregate retail investors net buy (sell) after negative (positive) returns. Kelley and Tetlock (2013) find that passive limit orders receive compensation from return reversals, but they do not find the same pattern for the more aggressive and more informed marketable orders, therefore leaving the liquidity provision of marketable orders unexplained. Boehmer, Jones, Zhang, and Zhang (2021) use the same TAQ dataset as this paper. They develop a method which is able to extract marketable orders from the TAQ data. They find that marketable orders of retail investors indeed provide liquidity, and retail order imbalances predict next period cross section of stock returns. However, they find that liquidity provision has no significant contribution to stock return predictability of order imbalance. Baker and Stein (2004) build a behavioral model to explain the liquidity provision of uninformed investors. They attribute liquidity provision by uninformed investors to their underreaction to the information contained in the order flow of other informed investors. Although they admit that the underlying behavioral mechanisms that might give rise to the underreaction can be many, in their model they assume that it is due to investors' overconfidence in their own private information. In this paper we find that when retail investors are able to obtain information from social media, they will largely reduce their liquidity provision. Our finding is consistent with the underreaction story, but it implies that the underreaction is due to retail investors' limited access to private information. Retail investors have only bounded rationality in learning from the trading of those investors who have private information, this makes them underreact to the information. When private information is released to them via social media in the form of readable articles with clear long or short recommendations, they can immediately trade on the private information, hence become less contrarian.

2.4 Informativeness of Short Sales

There is a vast literature showing that short sellers are informed traders. Theoretical work by Diamond and Verrecchia (1987) argue that because short sales constraints make liquidity traders have no use of the short sale proceeds, most short sellers are informed traders. Early empirical studies use monthly short interest to proxy the information in short sales, and find that when short interests are high, future returns are predictably low. Senchack and Starks (1993) find significant but small negative abnormal returns around short interest announcement date.Asquith, Pathak, and Ritter (2005) find that portfolios of stocks with high short interest (high demand of short sale) and low institutional ownership (low supply for short sale) underperform the market. More recent studies use daily short flow to proxy the information in short sales. Boehmer, Jones, and Zhang (2008) use proprietary NYSE order records related to short sales and find that heavily shorted stocks significantly underperform lightly shorted stocks. Among different account types, they find that institutional nonprogram short sales are the most informative. They also find that short flow data dominate short interest data in predicting future returns. Engelberg, Reed, and Ringgenberg (2012) use the transaction-level short sale data from TAQ Regulation SHO database combined with news release data from Dow Jones archive and find that the negative relation between short sales and future stock returns are doubled on news days and quadrupled on negative news days. Their findings suggest that a substantial portion of the information advantage of short sellers arises from their superior public information processing (interpreting) skills. Hu, Jones, Zhang, and Zhang (2021) use Cboe short-sale transaction data combined with data from social media platform Reddit for a short sample period (from January 2020 to February 2021) and find that when there is higher traffic in Reddit, the shorting flows become more informative in negatively predicting future returns. However, in their research design they use the interactions of two continuous variables (short sales × Reddit traffic volume), which makes it difficult to interpret their estimated coefficients. In this paper, we use Cboe short-sale transaction data combined with social media data from Seekingalpha, which has a longer sample period (from January 2008 to December 2019). We find that when there are articles published on SA, and the articles are commented by viewers of the article, the predictability of short sales on next day return will increase substantially. This result suggests that short sellers become more informed because of the SA articles. Interestingly, we find that short sellers trade against the noises of SA articles whereas retail investors trade in the same direction as the noises, which suggests that short sellers are superior in processing the information from social media.

2.5 Other Researches Using the SA Data Set

Lastly, this paper adds to the growing literature which uses SA data for their studies. Farrell, Green, Jame, and Markov (2018) find that an exogenous reduction of SA article coverage significantly results in an increase of market liquidity, which is measured by bid-ask spread or Amihud illiquidity measure. Campbell, DeAngelis, and Moon (2019) find that the position disclosure in SA articles increase the informativeness of the articles. Gomez, Heflin, Moon, and Warren (2020) show evidences that financial analysis on SA can reduce information asymmetry between less sophisticated investors and more sophisticated investors. First, they find a significant decline in bid-ask spreads (the proxy for information asymmetry) immediately following the publications of SA articles. Second, they find smaller spikes of bid-ask spreads at earnings announcement for quarters containing more SA articles about the firm. Drake, Moon, Twedt, and Warren (2022) find that the market reaction to the news in a sell-side analyst forecast is substantially reduced when preceded by the publications of SA articles. Shanthikumar, Wang, and Wu (2020) find that social media interaction moderates extremeness of SA comments and investor disagreement decreases significantly after SA articles, but not after analyst forecast days or high news days. Dim (2021) finds that there is heterogeneity of the ability of SA analysts in forming buying or selling beliefs. Only about 10% percent of SA analysts are skilled enough in producing economically meaningful abnormal returns. He also finds that SA analysts herd when they form beliefs. In this paper, we enhance the SA data with data of retail investor and short seller order flows, and highlight the implications of combining trading data with SA data for predicting equity prices.

3 DATA AND DESCRIPTIVE STATISTICS

3.1 Seeking Alpha

Seeking Alpha (SA) is one of the world's largest online investing communities. As of January 2021, the website had 10 million registered users and attracts over 17 million unique viewers every month. SA articles covers a broad range of stocks, ETFs and mutual funds, commodities and cryptocurrency, including thousands of stocks (such as small-caps) not analyzed elsewhere[1]. A large part of these articles are contributed by buy side investors and industry experts rather than sell side analysts[2]. Each month, about 10,000 investing ideas are published by more than 7000 contributors. Each article undergoes editorial review to ensure quality of the contents. Other interested investors can post their commentaries in response to an article. We develop a web-scraping algorithm to download all articles and comments that were published between 2005 and 2019. We then write programs to extract relevant information from the downloaded HTML and JSON files. Some of the articles have a "SA transcript" tag, They are not opinion (analysis) articles but earnings call transcripts. We exclude these articles from our study. Because the number of articles published on SA is small in the first two years of its foundation, we limit the sample period from 2007 to 2019.

SA editors tag one or more stock tickers to an article if it analyzes specific stocks instead of industry or macroeconomic conditions. We only use the single-ticker articles for our cross-sectional analysis. The single-ticker articles account for 92% of the articles that are tagged with stock tickers. Using the stock tickers and timestamps of single-tciker web pages, we construct a panel data of firm-day levels. For consistency with transaction data and return data, we define the SA articles and comments of date t as the ones that are published between the close hours of equity markets of previous day $(t-1)$ to the close hour of day t. Three groups of variables are extracted form the SA article and comment data:

1. $\mathrm{I}_{i,t}^{\text{Article}}$ and $\mathrm{I}_{i,t}^{\text{Comment}}$. These two binary variables indicate whether there is at least one article /

[1] https://seekingalpha.com/page/about_us

[2] https://en.wikipedia.org/wiki/Seeking_Alpha

comment published about stock i at date t.

2. $\log(\#)_{i,t}^{\text{Article}}$ and $\log(\#)_{i,t}^{\text{Comment}}$. These two variables are the log number of articles / comments published about stock i at date t, which is computed as logarithmic of $(1 + \#)$ of articles / comments.

3. $\text{NegSA}_{i,t}^{\text{Article}}$ and $\text{NegSA}_{i,t}^{\text{Comment}}$. These two variables measure the average negativeness of all the articles / comments of stock i at date t.

The negativeness of a document (either an article or a comment) is defined as the sum of normalized term frequency (TF) of negative words in a document. The list of negative words are compiled by Loughran and McDonald (2011), which are used extensively in literature in measuring tone of financial texts. We follow a standard procedure in NLP to compute TF of negative words (see Hapke, Howard, and Lane (2019)). First we tokenize the document into a list of words. Then we delete the uninformative tokens such as punctuation and stop words from the list. Lastly, we compute the normalized term frequency of a negative word as the number of times the word appears in the list divided by the number of unique words of the list (the length of the bag of words).

Following previous research, we use negativeness of a text to measure the tone of SA articles and comments. Tetlock, Saar-Tsechansky, and Macskassy (2008) show that one can interpret a low fraction of negative words as positive news. We don't use the Loughran and McDonald (2011) positive words list because positive words are often negated to convey negative feelings (see Chen, De, Hu, and Hwang (2014)).

In recent years, Seeking alpha phases in a new feature that labels the opinion articles as either "Very Bullish", "Bullish", "Neutral", "Bearish", or "Very Bearish". However, many of the articles published in earlier years have no such labels. We adopt the Support Vector Classifier (SVC) machine learning algorithm to label all the articles (See Dim (2021)). To reduce the imbalanced data problem, we collapse the belief labels to three classes, "bullish", "neutral", and "bearish", setting the "very bullish" and "bullish" to "bullish", the "very bearish" and "bearish" to "bearish"; the last label is "neutral". We thus extract the fourth group of variables to extend our analysis:

4. $\mathrm{I}_{i,t}^{\text{BearishArticles}}$ and $\mathrm{I}_{i,t}^{\text{BullishArticles}}$, which are two binary variables indicate whether stock i has at least one bearish article / bullish article at date t.

3.2 TAQ Retail Order Flows

Following the method by Boehmer, Jones, Zhang, and Zhang (2021), we identify transactions by retail investors from the Trades and Quotes (TAQ) database. The method is built on the fact that, due to regulatory restrictions in the U.S., retail order flow, but not institutional order flow, can receive price improvement, which is in small factions of a cent per share. We identify the trades with execution prices with a sub-penny portion between \$0.0001 and \$0.0040 as retail sells, and identify those with execution prices with a sub-penny portion between \$0.0061 and \$0.0099 as retail buys. We then compute the net sell (net buy) order flow by retail investors in terms of trading volume and number of transactions respectively.

$$\text{NetSell}^{\text{vol}} = \frac{\text{SellVol} - \text{BuyVol}}{\text{TotalVol}}, \qquad \text{NetSell}^{\text{trans}} = \frac{\text{SellTrans} - \text{BuyTrans}}{\text{TotalTrans}}$$

$$\text{NetBuy}^{\text{vol}} = \frac{\text{BuyVol} - \text{SellVol}}{\text{TotalVol}}, \qquad \text{NetBuy}^{\text{trans}} = \frac{\text{BuyTrans} - \text{SellTrans}}{\text{TotalTrans}}$$

3.3 CBOE Short Sales

Cboe is currently one of the largest U.S. equities market operators. It operates four U.S. equities exchanges, the BZX Exchange, the BYX Exchange, EDGA Exchange, and EDGX Exchange[3]. To help increase market transparency, the Cboe U.S. Equities Exchanges make short sale information publicly available.[4] The same day short sale transactions and volume data are published and free to download after the close of the market. Hu, Jones, Zhang, and Zhang (2021) find that Cboe accounts for about 20% of on-exchange shorting activity on average.

For each day , we sum the short volume of a stock in the four exchanges, then compute the

[3] https://www.cboe.com/us/equities/overview/

[4] https://www.cboe.com/us/equities/market_statistics/short_sale/

short sale variable as proposed by HJZZ:

$$\mathrm{SS}_{i,t} = \frac{\text{Daily Cboe short volume}_{it}}{\text{Total CRSP trading volume}_{it}}$$

Cboe short sale data is available from 2008.

Earlier studies use monthly short interest data to proxy information of short sales. Cboe data has two advantages over the traditional short interest data. First, Cboe short sale data are much finer than the monthly short interest. Second, short interest data can only record the uncovered positions but cannot capture the short-lived "in-and-out" shorting that could be prevalent.

3.4 Informational Variables

We use a group of variables to proxy information of stock fundamentals or to measure information environment of the stocks. These variables are computed using data from several sources.

3.4.1 Earnings Level and Earnings Growth

We compute the price adjusted earnings and earnings growth using quarterly fundamentals dataset of Compustat. We define price adjusted earnings as the ratio of earnings per share and price at the end of quarter: $\text{AdjEarings}_q = \text{EPS}_q/\text{Price}_q$. We define one year earnings growth as $(\text{EPS}_q - \text{EPS}_{q-4})/\text{EPS}_{q-4}$ and one quarter earnings growth as $(\text{EPS}_q - \text{EPS}_{q-1})/\text{EPS}_{q-1}$.

3.4.2 IBES Analyst Forecast

IBES analyst forecast revisions variables captures the major changes of the expectations of sell side analysts about the firms' earnings growths. The forecast revisions are of different horizons which reflect different aspects of the changes of firms' fundamental values. The specific definitions of the group of revision variables (Numup and Numdown) are listed in Table C.1. IBES analyst coverage is the number of analysts who give the forecast of EPS of a stock in a given month.

3.4.3 IBES Recommendations

IBES provides buy / sell recommendations of the stocks. The scale is: 1. Strong buy; 2. Buy; 3. Hold; 4. Underperform; 5. Sell. We use the average recommendations as well as the number of recommendations up and down from the IBES summary statistics data set.

3.4.4 RavenPack News

We use the Event Sentiment Score (ESS) of Ravenpack to measure sentiment across a comprehensive set of cash flow relevant news about a stock on a given day. We use the Aggregate Event Volume (AEV) to measure the volume of the cash flow news.

3.5 Returns and Abnormal Returns

We use the daily return from the CRSP daily stock files. We also compute the buy and hold abnormal returns using the DGTW characteristic-based benchmark method (see Daniel, Grinblatt, Titman, and Wermers (1997)):

$$\mathrm{AR}_{i,p} = \mathrm{R}_{i,p} - \mathrm{R}_{i,p}^{\mathrm{Benchmark}}$$

where $\mathrm{R}_{i,p}$ is the cumulative return of stock i during period p, and $\mathrm{R}_{i,p}^{\mathrm{Benchmark}}$ is the cumulative return of the benchmark portfolio of stock during the same period. The benchmark portfolio return is the value-weighted returns of stocks matched on size, book-to-market, and momentum in a sorting. The size of a stock is the market capitalization of the stock in prior year, which is computed as the product of price and shares outstanding of the last trading day of prior year. The momentum of a stock is the cumulative return of the stock in prior 12 months (skip the most recent month). The book-to-market is the ratio of book value and market capitalization of prior year. The book value is the common/ordinary equity (CEQ) variable from Compustat annual fundamentals dataset. We also compute the volatility of a stock at date as the variance of daily returns in the preceding (following) 21 trading days when it is used as a RHS (LHS) variable.

3.6 Descriptive Statistics

Table B.1 reports the descriptive statistics of the major variables used in this paper. The numbers in Panel A and Panel B are computed as the time-series averages of the daily cross-sectional statistics. Panel A reports the statistics of individual variables, which includes the Mean, standard deviation, Min, Max, 25%, Median, and 75%. From Panel A we can see that in the sample the daily average of the gross returns of the stocks is about 6bps, whereas that of the abnormal returns is close to 0, which is due to the adjustment of benchmark returns. The average negativeness of SA articles (comments) is 3.3% (2%). On average retail investors slightly net sell the individual stocks (3% in trading volume and 2% in number of transactions). The average of the adjusted short sale measure SS is about 8%. The average number of IBES analyst that covers the individual stock is around 6.4, and the average number of upward (downward) revision is around 1.2 (1.8). The average Ravenpack news sentiment (ESS) is around 0.53, and the logarithmic of news volume is around 1.82. The quarterly earnings growth of individual stocks has an average about 0.2%, but has a big dispersion (the standard deviation of which is 1.18). From the column of standard deviation we can observe that each individual variable has enough cross-sectional variations for later regression analysis.

From the correlation table of panel B, we can see that in cross section the negativeness of SA articles and comments are positively correlated with the flows of retail-investor net sells, the flows of short sales, number of analyst upward revisions, and the volume of Ravenpack news, but negatively correlated with returns, abnormal returns, the number of analyst estimates, the number of analyst upward revisions, the positiveness of Ravenpack news, and earnings growth. These results show that the tone of SA articles is informative in conveying fundamental information of individual stocks. From the table, we further notice that although the negativeness of SA articles (comments) is negatively correlated with the positiveness of Ravenpack news, the correlation is relatively low, which implies that social media contribute incremental information to the markets beyond that of conventional news channels.

Panel C of Table B.1 reports the annual coverage by SA articles of the individual stocks in the sample. In Column 3 are the total number of individual stocks in the sample included in the sample each year, in column 4 are the number of individual stocks covered by SA articles. As the SA website grows, the SA article coverage ratio increases from less than one third in 2007 to about sixty percent in 2019.

4 RETAIL INVESTORS

4.1 Social Media Tone and Order Imbalances of Retail Investors

In this section we provide evidence that social media tone are informative about retail order flows. We regress retail order imbalances on SA article and comment negativeness. The observations are on a firm-day level. For an observation to be included in the sample of regression, we require it to have at least one SA article or comment in the firm-day combination. The formal equation is,

$$\text{NetSell}_{i,s} = \beta_1 \text{NegSA}^{\text{Article}}_{i,t} + \beta_2 \text{NegSA}^{\text{Comment}}_{i,t} + \beta_3' \text{Controls}_{i,t} + m_t + f_i + \varepsilon_{i,s} \tag{1}$$

where $s \in \{t, t+1\}$. Thus we estimate the influence of social media opinion of date t on both contemporaneous (t) order flows and that of next trading day ($t+1$). The list of control variables are abnormal return of current date AR$_t$, abnormal return of prior five trading days AR$_{[t-5,t-1]}$, momentum of prior year, and volatility of prior month. We also include year-month fixed effect m_t and firm fixed effect f_i in the regression.

The results of regression are reported in Table B.2. For columns 1, 2, and 3, the LHS variable is computed using trading volume of retail investors; for columns 4, 5, and 6, the LHS variable is computed using the number of transactions by retail investors. The results show that, as tone of SA articles and comments of a specific stock is more negative, the aggregate retail investors tend to net sell the stock more today and on next trading day. The effect is statistically significant ($t = 4.68$ for article and $t = 2.39$ for comment in column 3) but economically small. One standard deviation increase of SA article (comment) negativeness will increase retail net sells by 27 bps (10 bps). Our rationale is that only part of the retail orders are driven by social media effect, other orders reflect the tradings for liquidity needs, or are plainly ignorant of the news on SA due to limited attention. Hence the effect on aggregate level is small.

4.2 Social Media Talks That Most Predict Retail Order Imbalances

In the previous section we find that the talks on social media are informative about real world retail order flows. In this section we want to explore what types of talks in SA articles are most relevant about next day retail order imbalances. Our method proceeds as follows. We label each SA article of a specific stock by the direction of the next day retail order imbalance of the stock. Thus each labeled article belongs to either of the two groups: the net buys and the net sells. We use the TF-IDF of the words (bigrams) of the article as features. We then use the labeled sample to train a Support Vector classifier. We then use the weight vector of the SVC as a measure of the words' relevance on the classification. The negative (positive) weights contribute to the negative (positive) classification. In Figure A.1 we report the top 50 words (bigrams) contribute most to the two classes.

We can see from Figure A.1 that the terms related to retail net buys (such as `complete`, `surge`, `synergy`, `competitive_advantage`, `efficiency`, `potential`, etc.) tend to be more positive and the terms related to retail net sells (such as `division`, `restructuring`, `decline`, `cloud`, `bad`, `problem`, etc.) tend to be more negative. This result provides intuitive evidence that retail investors do follow the analysis of SA articles in making buy and sell decisions.

4.3 Impact of Social Media on Liquidity Provision of Retail Investors

Prior studies (Kaniel, Saar, and Titman (2008), Kelley and Tetlock (2013), Barrot, Kaniel, and Sraer (2016), Boehmer, Jones, Zhang, and Zhang (2021)) find that retail investors tend to take reversal (contrarian) strategies in buying and selling stocks, which make them take opposite positions to rest of the market. Therefore retail investors act as liquidity providers for the market. Kelley and Tetlock (2013) find that the contrarian strategy of passive limit orders benefit from the reversal of temporary price movement caused by distortions of institutional investors. However, it cannot be used to explain the liquidity provision also observed in the more aggressive marketable orders, which they find do not benefit from price reversals. We conjecture that one reason for retail orders to be contrarian is that as unsophisticated investors they underreact to information contained in the order flow of informed traders (Baker and Stein (2004) hold a similar view in their theoretical work). SA articles, on the other hand, can provide retail investors the private information in more readable forms and with more actionable buying and selling recommendations. It will help retail investors to react to the information more timely and therefore largely reduce liquidity provision. We test this conjecture by the following estimates:

$$\text{NetBuy}_{i,t+1} = \sum_{p} \left(\beta_{1,p}\text{AR}_{i,p} + \beta_{2,p}\text{AR}_{i,p} \times \text{I}_{i,p}^{\text{Article}} + \beta_{3,p}\text{I}_{i,p}^{\text{Article}}\right) + m_t + f_i + \varepsilon_{i,t+1} \tag{2}$$

where $p \in \{t, [t-5, t-1], [t-26, t-6]\}$ represents one of the three periods before date $t+1$. $\text{AR}_{i,p}$ is the buy and hold abnormal return of stock during period p. $\text{I}_{i,p}^{\text{Article}}$ is a dummy variable which equals 1 when there are SA article coverage about stock i during period p, and equals 0 when there are no SA article coverage. m_t is year-month fixed effect, f_i is firm fixed effect. Because retail investors take the contrarian strategies, we expect that the coefficient $\beta_{1,p}$ is negative, i.e., when previous returns of stock i are negative, retail investors tend to buy the stock, and when previous returns of stock i are positive, they tend to sell the stock. Because easy information from social media make retail trades less contrarian, we expect that $\beta_{2,p}$ is positive.

The results of the regression are reported in columns 1, 2 and 4, 5 of Table B.3. Column 1 and 2 use trading volume to calculate LHS variables. Column 4 and 5 use number of transactions to calculate LHS variables. Column 1 and 4 replicate the findings of Barrot, Kaniel, and Sraer (2016) and Boehmer, Jones, Zhang, and Zhang (2021) that marketable order imbalances are contrarian. Column 2 and 5 show the effect of SA article on liquidity provision of retail investors. We can see from the table that indeed $\beta_{1,p}$ is negative and $\beta_{2,p}$ is positive. The effect of social media on retail trading strategies is economically large. As shown in column 2, the contrarian trading at date t by retail traders for the stocks that have SA article coverage during period $[t-5, t-1]$ ($[t-26, t-6]$) are 51% (33%) smaller than the stocks that do not. The results in this section suggest that SA articles indeed facilitate retail investors in acquiring and interpreting information, so that they become more fast in reacting to the private information that they had difficulty to learn from informed investors' order flows or market price signals before.

4.3.1 Investor Attention

One potential endogeneity problem of the analysis in the previous section is that the effect we find could just be from other news sources that published at approximately the same time as the SA articles. To rule out this alternative explanation, we modify the previous regression as:

$$\begin{aligned}\mathrm{NetBuy}_{i,t+1} = \sum_{p} \Big(&\beta_{1,p}\mathrm{AR}_{i,p} + \beta_{2,p}\mathrm{AR}_{i,p} \times \mathrm{I}^{\mathrm{AC}}_{i,p} + \beta_{3,p}\mathrm{AR}_{i,p} \times \mathrm{I}^{\mathrm{A\overline{C}}}_{i,p} + \beta_{4,p}\mathrm{I}^{\mathrm{AC}}_{i,p} + \beta_{5,p}\mathrm{I}^{\mathrm{A\overline{C}}}_{i,p}\Big) \\ &+ m_t + f_i + \varepsilon_{i,t+1}\end{aligned} \tag{3}$$

where $\mathrm{I}^{\mathrm{AC}}_{i,p}$ is the dummy variable indicating that stock had SA articles during period p and these articles received comments; $\mathrm{I}^{\mathrm{A\overline{C}}}_{i,p}$ is the dummy variable indicating that stock i had SA articles during period p, but these articles did not receive comments. The former (latter) dummy variable represents the scenario where the SA articles attracted (did not attract) investor attention.

The results of the regression are reported in columns 3 and 6 of Table B.3, which show that only the SA articles that attracted enough investor attention can have a significant negative effect on aggregate retail investors' liquidity provision. This result demonstrates that what we find is not merely effect of news from external sources, but from investors adopting opinions of SA articles.

4.4 Impact of Social Media on Return Predictability of Retail Orders

In this section we study how social media influence the informativeness of retail orders in predicting future stock returns. Recent researches (Kelley and Tetlock (2013), Boehmer, Jones, Zhang, and Zhang (2021)) show that retail order imbalances positively predict future returns, which demonstrates that retail investors are not just noise traders as demonstrated by earlier research (See Odean (1999), Barber and Odean (2000), Grinblatt and Keloharju (2001), Hvidkjaer (2008)). After all, retail investors are not subject to the agency problems, career concerns, or liquidity constraints that can hurt institutional managers' performance, which makes them have incentives to trade on novel cash flow information and ultimately profit from the trading. Our analysis in the previous section shows that information from social media make retail investors move away from contrarian tradings. We conjecture that access to social media information will help retail investors realize higher future returns. We test the conjecture using the following estimates:

$$\text{AR}_{t+1} = \beta_1 \text{NetBuy}_{i,t} + \beta_2 \text{NetBuy}_{i,t} \times \text{I}_{i,t}^{\text{Article}} + \beta_3 \text{I}_{i,t}^{\text{Article}} + \sum_p \text{AR}_p + m_t + f_i + \varepsilon_{i,t+1} \tag{4}$$

where $\text{I}_{i,t}^{\text{Article}}$ is a dummy which equals 1 when there are SA articles about stock i on date t. AR_ps are abnormal returns of previous periods included as control variables. m_t and f_i are year-month fixed effect and firm fixed effect respectively.

The results of the regression are reported in column 2 of Table B.4. One standard deviation increase of net buys of the stocks that have no SA articles will increase the next day abnormal returns by 2bps, whereas one standard deviation increase of the net buys of the stocks that have SA articles will increase the next day abnormal returns by 7bps. The 5 bps difference is both statistically significant ($t = 3.21$) and economically large. Same as the previous section, we further decompose the dummy variable $\text{I}_{i,t}^{\text{Article}}$ to two dummy variables $\text{I}_{i,t}^{\text{AC}}$ and $\text{I}_{i,t}^{\text{A}\overline{\text{C}}}$, which represent that stock i receives commented SA articles and non-commented SA articles at date t respectively. The regression result is in column 3 of Table B.4. We find that the coefficient of $\text{NetBuy}_{it} \times \text{I}_{i,t}^{\text{AC}}$ is 6bps and is significant at the 5% level, and the coefficient of $\text{NetBuy}_{it} \times \text{I}_{i,t}^{\text{A}\overline{\text{C}}}$ is 4bps and is marginally significant at the 10% level. However, the null hypothesis of equality of the two coefficients cannot be rejected (p value is 0.58). This result suggests that attention of the SA articles from retail investors affects diffusion of the information in articles to retail investors.

We further study whether the trading profits of retail investors are from they buying stocks that are covered by bullish SA articles or from selling stocks that are covered by bearish SA articles. Because the variable NetBuy_{it} can be either buying (when it is greater than 0) or selling (when it is less than 0), we further decompose it into $\max(\text{NetBuy}_{it}, 0)$ and $\max(\text{NetSell}_{it}, 0)$. Column 4 of Table B.4 shows that the coefficient of the interaction term $\max(\text{NetBuy}_{it}, 0) \times \text{I}_{i,t}^{\text{BullishArticle}}$ is

31 bps and is significant ($t = 4.05$), but the coefficient of the term $\max(\text{NetSell}_{it}, 0) \times \text{I}_{i,t}^{\text{BearishArticle}}$ is insignificant. The result indicates that trading profits are more from buying on the bullish news than selling on the bearish news on social media.

4.5 Portfolio Analysis

Prior research that uses opinions of SA articles and comments for constructing investment portfolios achieves only moderately better performance than the market portfolio (See Chen, De, Hu, and Hwang (2014) and Dim (2021)). The reason could be that there are heterogeneity of SA authors in forming correct investment opinions, which makes signals of social media too noisy. To better utilize the wisdom of crowds, one can combine the signals of investor talks on social media and that of their walks in real world transactions. This is the methodology we take in this section as well as in section 5.5.

In this section we construct two zero-cost portfolios mimicking trades of retail investors. The difference of the two portfolios depends on whether they have access to information of social media. At each date t, both of the portfolios are long the stocks that aggregate retail investors net buys ($\text{NetBuy}_{it} > 0$) and are short the stocks that aggregate retail investor net sells ($\text{NetBuy}_{it} < 0$). For the first portfolio, we further require that each stock i in the portfolio (either in the long leg or short leg) need to have SA articles posted on date t ($\text{I}_{it}^{\text{Article}} = 1$); whereas for the second portfolio, we require that none of the stocks have SA articles posted on date t ($\text{I}_{it}^{\text{Article}} = 0$). We then calculate one trading day ahead equal-weighted returns of the two portfolios $\overline{\text{R}}_{t+1}$. Hence the two portfolios are rebalanced every trading day. For convenience, we call the first portfolio informed portfolio, and the second uninformed portfolio.

Figure A.3 shows the monthly evolution of $1 invested in the long leg and short leg of the informed portfolio (*retail net buy, SA* and *retail net sell, SA* in the figure respectively), the long leg and short leg of the uninformed portfolio (*retail net buy, no SA* and *retail net sell, no SA* in the figure respectively), and the market portfolio (Market in the figure). The two dashed lines record the performance of the two long legs. During the sample period from January 2010 to December

2018, the value of the long leg of the informed portfolio (*retail net buy, SA*) increases from \$1 to its peak \$11.6 in 2018, but has a big pull back (28%) in the same year, the value of the long leg of the uninformed portfolio (*retail net buy, no SA*) increases from \$1 to its peak \$6.6 in 2018, and also has a big pull back (20%) in the same year. Both of the short legs of the informed and uninformed portfolios (in solid line in the figure) underperform the market, but both realize a positive return. In the attached table of Figure A.3, we also list the Sharpe ratios of the five portfolios computed using their daily returns. The Sharpe ratios of the long legs of both the informed (0.08) and uninformed (0.07) exceed that of the market portfolio (0.05), but that of the informed portfolio is higher. The Sharpe ratios of the short legs of the informed and uninformed portfolios are similar (0.03), which are below that of market portfolio.

The above analysis suggests that trading strategy combining information of social media and retail order flows has better performance than that using retail order flows information alone.

4.5.1 Alphas and Risk Loadings

We further regress the daily time series of the two portfolios on major risk factors. We first run Fama-French five-factor regression,

$$\mathrm{R}_t - \mathrm{rf}_t = \alpha + \beta_1 \mathrm{Mkt\text{-}rf}_t + \beta_2 \mathrm{SMB}_t + \beta_3 \mathrm{HML}_t + \beta_4 \mathrm{CMA}_t + \beta_5 \mathrm{RMW}_t + \varepsilon_t \tag{5}$$

and add three more factors in the second estimation:

$$\begin{aligned}\mathrm{R}_t - \mathrm{rf}_t = \alpha &+ \beta_1 \mathrm{Mkt\text{-}rf}_t + \beta_2 \mathrm{SMB}_t + \beta_3 \mathrm{HML}_t + \beta_4 \mathrm{CMA}_t + \beta_5 \mathrm{RMW}_t \\ &+ \beta_6 \mathrm{MOM}_t + \beta_7 \mathrm{STRev}_t + \beta_8 \mathrm{LTRev}_t + \varepsilon_t\end{aligned} \tag{6}$$

Table B.5 reports the risk loadings and annualized alphas of the informed and uninformed portfolios. Column 1 and 2 show that the informed portfolio has risk loading on the market factor. To manage the risk of the informed portfolio, one can sell forward market index to hedge its exposure to market factor. Column 3 and 4 show that the uninformed portfolio has positive loading on size factor, whereas column 1 and 2 show that informed portfolio has no positive loading on size factor. This result suggests that uninformed retail tradings tend to buy small stocks whereas informed retail tradings do not. Both of the informed portfolio and uninformed portfolio have no loading on short-term and long-term reversal factors. This result is consistent with that of Kelley and Tetlock (2013), who find that, unlike passive limit orders, more aggressive market orders do not gain profits from future return reversals.

5 SHORT SELLERS

5.1 Social Media Tone and Short Sale Flows

In this section we provide evidence that social media tone is informative in predicting short sale flows. In the following estimations, we regress the short sales of stocks of the contemporaneous (t) and next trading day ($t + 1$) on average negative tone of SA articles and comments of the stocks of day t. The observations are on a firm-day level. For an observation to be included in the sample of regression, we require it to have at least one SA article or comment for the firm-day combination. The regression equation is:

$$\text{SS}_{i,s} = \beta_1 \text{NegSA}_{i,t}^{\text{Article}} + \beta_2 \text{NegSA}_{i,t}^{\text{Comment}} + \beta_3' \text{Controls}_{i,t} + m_t + f_i + \varepsilon_{i,s} \tag{7}$$

where $s \in \{t, t{+}1\}$. The list of control variables are abnormal return of current date AR_t, abnormal return of prior five trading days $\text{AR}_{[t-5,t-1]}$, and volatility of prior month. We also include year-month fixed effect m_t and firm fixed effect f_i in the regression.

The results of the estimation is in Table B.6. Column 1, 2, 4, 5 show that the positive correlation between negativeness of SA articles / comments and short sales are significant. It is possible that short sales predate the release of negative news on social media, and short sales are persistent due to drift of stock prices, thus one could falsely observe prediction of social media tone on future short sales. For instance, it is possible that the short sellers could receive and trade on new information earlier than less sophisticated investors, or it is even possible that short sellers first sell a stock short on the market, then post negative opinions about the stock on social media. To alleviate this endogeneity issue, in column 3 and 6, where the LHS vriable is $\text{SS}_{i,t+1}$, we add $\text{SS}_{i,t}$ on the right hand side as control variable. We can see that indeed short sale is highly persistent ($t = 64.49$ in column 3). The predictability of the tone of SA articles on short sales become less, but still significant after adding the lag short sales as control variable, whereas the predictability of the tone of SA comments on short sale become statistically insignificant.

5.2 Social Media Noises and Short Sale Flows

The messages on social media do not always convey the correct information about the stock, they could also merely reflect investment sentiment unrelated to firm fundamentals. One interesting question to ask is how do the short sellers react to the irrational sentiment, or the noises of the market that is captured by social media? Do they trade against the noises, or do they trade in the same direction as market sentiment? As sophisticated investors, they can either trade against social media sentiment when they expect that a reversal of stock prices will happen shortly, or they do nothing because of fear of potential short squeezes, or they can jump on the bandwagon and trade in the same direction as the noise traders if they believe that the trend of present price movement will continue.

To answer the question, we design a two step analysis. First, we regress the tone of social media articles and comments on the variables that proxy information about the fundamental values of the stock at the time when SA articles (comments) are published. We then use the residuals of the regression as a proxy of the opinions of noise traders and to see its relation with short sales in the second step.

Our proxy variables of information include the level and growth of earnings (at announcement), revisions of IBES analyst forecast, mean and revisions of IBES analyst recommendation, and Ravenpack news sentiment. We require these news are released near the publishing date of SA articles / comments. We include the news released both before and after the publishing of SA articles/comments to accommodate the fact that social media analysts can know some new information before it is released to the public.

Table B.7 reports the estimates of the pooled regression of the first step. We can see from the results that news about firm fundamental values can only account for a small fraction of the variation of the social media opinion.

We define the residuals from the regressions in Table B.7 as new variables: *Abnormal Negativeness of SA Articles (comments)*, denoted as $\text{AbNegSA}_{i,t}^{\text{Article}}$ ($\text{AbNegSA}_{i,t}^{\text{Comment}}$). The estimate equation for the second step is:

$$\text{SS}_{i,s} = \beta_1 \text{AbNegSA}_{i,t}^{\text{Article}} + \beta_2 \text{AbNegSA}_{i,t}^{\text{Comment}} + \beta_3' \text{Controls}_{i,t} + m_t + f_i + \varepsilon_{i,s} \tag{8}$$

The results of the regressions are reported in Table B.8. The statistically significant negative coefficient of the variable $\text{AbNegSA}_{i,t}^{\text{Article}}$ shows that as the sentiment of noise traders reflected in SA articles is more positive, the sophisticated investors tend to sell the stock short more. In other words, short sellers tend to trade against the noises. The estimate of β_2 is not significant, it could be because of the uncertainty in evaluating the divergent opinions of commenters on the SA platform. Compare the results in Table B.6 and Table B.8, we demonstrate that short sellers can distinguish information and noise. In contrast, results of Table B.2 and Table C.2 suggest that retail investors cannot distinguish information and noise, because they tend to buy a stock when market sentiment unrelated to stock fundamental values is more positive and tend to sell it when market sentiment is more negative.

5.3 Social Media Talks That Most Predict Short Sale Levels

In this section we want to explore what types of talks in SA articles are most relevant about level of next day short sales. We use the same method as section 4.2. For each trading day we split the stocks into lightly shorted group and heavily shorted group by the midpoint of next day short sale: $\text{SS}_{i,t+1}$. Then the classification of lightly or heavily shorted is the label of an article of a specific stock and TF-IDF of the article is the feature. We then train the SVC model as section 3.2 and extract the most relevant words for the lightly shorted and heavily shorted groups. In Figure A.2 we report the top 50 words (bigrams) most relevant to the two groups.

As shown in Figure A.2, not like the relation between retail orders and social media talks, the relation between short sale level and social media talks is not obvious. Interestingly, the bigram `disclosure_long` is strongly related to a heavy short sale next day. This bigram, restored to sentence by adding back the stop words, is actually the disclosure statement at the end of the analytical article: “`Disclosure:  I am (we are) long XXX stock...`”. It provides an interesting anecdotal evidence that short sellers trade against the opinions of SA articles. The short sellers perhaps take the disclosed long position of SA authors as an valuable shorting opportunity.

5.4 Impact of Social Media on Return Predictability of Short Sales

One of the most solid results in empirical finance literature is the negative predictability of short sales on future stock returns. In this section, we show that information of social media can further improve return predictability of short sales. The estimate equation is:

$$\mathrm{AR}_{t+1} = \beta_1 \mathrm{SS}_{i,t} + \beta_2 \mathrm{SS}_{i,t} \times \mathrm{I}^{\text{Article}}_{i,t} + \beta_3 \mathrm{I}^{\text{Article}}_{i,t} + \sum_p \mathrm{AR}_p + m_t + f_i + \varepsilon_{i,t+1} \tag{9}$$

The coefficient of interest is β_2, which measures the effect of interaction of short sale and indicator of coverage of SA artciels on returns of next trading day.

The results of the regression are reported in column 2 of Table B.9. The estimated coefficient $\hat{\beta}_1$ indicates that one standard deviation increase of short sale predicts 2 bps decrease of next trading day return when there is no coverage of SA articles. The estimated coefficient $\hat{\beta}_2$ indicates that when there are articles about a stock posted on SA, the predicted return decrease is 5 bps higher.

Column 3 of Table B.9 shows the impact of commented articles $\mathrm{SS}_{i,t} \times \mathrm{I}^{\mathrm{AC}}_{i,t}$ and non-commented articles $\mathrm{SS}_{i,t} \times \mathrm{I}^{\mathrm{A}\overline{\mathrm{C}}}_{i,t}$. We can see the former has a significant effect of 6 bps ($t = -3.34$), whereas the latter has a marginally significant effect of 4 bps ($t = -1.84$). However, the null hypothesis that the two effects are equal cannot be rejected ($p = 0.51$).

The results of Column 4 of Table B.9 show that short sellers obtain information more from the bullish articles than from the bearish articles.

5.5 Portfolio Analysis

In this section we build two zero-cost portfolios based on short sale flows. At each trading day t, we divide the stocks into two groups by the midpoint of variable $\mathrm{SS}_{i,t}$. To construct the zero-cost portfolios, it is long the stocks in the low $\mathrm{SS}_{i,t}$ group and is short the stocks in the high $\mathrm{SS}_{i,t}$ group. For the informed portfolio we require that each stock in the portfolio have SA articles published at date t ($\mathrm{I}_{it}^{\text{Article}} = 1$), whereas for the uninformed portfolio, we require that none of the stocks have SA articles posted on date t ($\mathrm{I}_{it}^{\text{Article}} = 0$). We then calculate the equal-weighted returns of the two portfolios in the next trading day $\overline{\mathrm{R}}_{t+1}$. Hence the two portfolios are rebalanced every trading day.

Figure A.4 presents the monthly evolution of $1 investment in the long and short legs of the informed portfolio (*Lightly shorted, SA* and *Heavily shorted, SA* in the figure respectively), the long and short legs of the uninformed portfolio (*Lightly shorted, no SA* and *Heavily shorted, no SA* in the figure respectively), and the market portfolio (*Market* in the figure). The two dashed lines record the performance of the two long legs. During the sample period from January 2008 to December 2019, the value of the long leg of the informed portfolio (*Lightly shorted, SA*) increases from $1 to around $58, whereas the value of the long leg of the uninformed portfolio (*Lightly shorted, no SA*) only increases from $1 to around $12 dollars. Both of the short legs of the informed and uninformed portfolios (*Heavily shorted, SA* and *Heavily shorted, no SA* respectively, in solid line in the figure) slightly underperform the market, but both realize a positive return. In the attached table of Figure 4, we also list the Sharpe ratios of the five portfolios computed using their daily returns. We can see that the Sharpe ratios of the long legs of both the informed (0.08) and uninformed (0.06) are higher than that of the market portfolio (0.03), but that of the informed portfolio is even higher. The Sharpe ratios of the short legs of the informed and uninformed portfolios are similar (0.02), which are below that of market portfolio.

One observation from the results is that the profits of the both the informed and uninformed portfolios are from long leg but not short leg. This is consistent with the finding of Boehmer, Jones, and Zhang (2008), that "short sellers are particularly good at avoiding shorting undervalued stocks, ... but are not necessarily identifying stocks that are overvalued". "This suggests that it is better to think of short sellers as keeping price in line rather than bring prices back into line". Social media coverage also only improves profitability of the portfolio by avoiding undervalued stocks rather than identifying overvalued stocks.

5.5.1 Alphas and Risk Loadings

We regress daily returns of two "short seller" portfolios on Fama-French 5 factors and extended 8 factors. The results are reported in Table B.10. The informed portfolio realizes an annualized alpha around 38%, whereas the uninformed portfolio 14%. The informed portfolio only has a marginally significant negative loading on momentum factor (previous 11 days), whereas the uninformed portfolios have loadings on all factors except the market factor. The informed portfolio which combines the information of short sale flows and social media coverage is promising. One advantage of the strategy is that the daily data of SA article publications and Cboe short sales needed to construct the portfolio are all public and readily available after the closing of market each trading day. However, we acknowledge the potential challenges in implementing the strategy in real world, including the transaction cost of daily rebalancing, the probable bias from bid-ask bounce, etc.

6 INFORMATIONAL EFFICIENCY

In this section we evaluate the effect of social media on informational efficiency of stock prices. Our findings in the previous sections suggest that social media coverage makes both retail trading and short selling more informed. With more informed tradings, we conjecture that this will make information be impounded into stock prices more quickly.

We adopt the price-delay measure introduced by Hou and Moskowitz (2005) and Boehmer and Wu (2013), which estimates how quickly prices incorporate public information. The market return is employed as the relevant news to which the stocks respond. At the end of each month t, we run a regression of each stock's daily returns on contemporaneous and five days lagged returns on the market portfolio.

$$R_{id} = \alpha_{it} + \beta_{it} R_{m,d} + \sum_{n=1}^{5} \delta_{it}^{(-n)} R_{m,d-n} + \varepsilon_{id}$$

where R_{id} is the return of stock i on trading day d in month t, and $R_{m,d}$ is the return of the CRSP value-weighted market index on day d. If the stock responds immediately and accurately to market news, then β_{it} will be significantly different from 0, but none of the $\delta_{it}^{(-n)}$s will differ from 0. But if stock i's price responds market information with a delay, then some of the $\delta_{it}^{(-n)}$s will be significantly different from 0.

The first measure D_1 is defined as 1 minus the ratio of the R-squared of the restricted regression ($\delta_{it}^{(-n)} = 0$) and the unrestricted R-squared:

$$D_{1,it} = 1 - \frac{R^2_{\delta_{it}^{(-n)}=0, \forall n \in [1,5]}}{R_t^2}$$

The higher D_1 is, the more the contemporaneous returns are explained by the lagged market returns, hence the strong the delay in response to public news. To give more weight to the longer lags and also to consider the precision of the estimates, we also include D_2 and D_3.

$$D_{2,it} = \frac{\sum_{n=1}^{5} n \left|\delta_{it}^{(-n)}\right|}{|\beta_{it}| + \sum_{n=1}^{5} n \left|\delta_{it}^{(-n)}\right|}, \quad D_{3,it} = \frac{\sum_{n=1}^{5} n \frac{\left|\delta_{it}^{(-n)}\right|}{\mathrm{se}\left(\delta_{it}^{(-n)}\right)}}{\frac{|\beta_{it}|}{\mathrm{se}(\beta_{it})} + \sum_{n=1}^{5} n \frac{\left|\delta_{it}^{(-n)}\right|}{\mathrm{se}\left(\delta_{it}^{(-n)}\right)}}$$

We thus get a firm-month panel data of D_1, D_2, and D_3, and use them as LHS variables in the following regressions:

$$\mathrm{D}_{j,it} = \beta_1 \log(\#)_{it}^{\text{Article}} + \beta_2 \log(\#)_{it}^{\text{Comment}} + \beta_3 D_{j,i,t-1} + \beta_4' \text{Controls}_{it} + m_t + f_i + \varepsilon_{i,t} \quad (10)$$

where $j \in \{1, 2, 3\}$ index the three price-delay measures. $\log(\#)_{it}^{\text{Article}}$ and $\log(\#)_{it}^{\text{Comment}}$ are the log number of articles and comments posted about stock i in month t.[5] The list of control variables are firm size, book-to-market ratio, momentum, analyst coverage, and log value of Ravenpack aggregate event volume ($\log$(AEV)).

The results of the regressions are reported in Table B.11. Columns 1, 3, 5 are the regressions without controls, columns 2, 4, 6 are the regressions with controls. The results show that on average the SA articles improve efficiency because one percentage increase of article posts decrease the price delay by about 2 percentage point; and on average the SA comments reduce efficiency because one percentage increase of comment posts increase the price delay by about 1.5 percentage point.

As a robustness check, we also change the RHS variables to $\mathrm{I}_{i,t}^{\text{Article}}$ and $\mathrm{I}_{i,t}^{\text{Comment}}$. The results of the estimation are reported in Table C.4. We can see that results of Table C.4 are consistent with that of Table B.11.

Volatility of stock returns is another proxy for informational efficiency because the stock prices become more volatile as there are more noise tradings. In Table C.5, we report the results of Fama-MacBeth regressions (Fama and MacBeth, 1973) of which return volatility is the RHS variable.

[5] To be precise, if the first and last trading day of month t are d_1 and d_2 respectively, the number of articles and comments in month t is computed as the number of articles and comments published between $d_1 - 5$ and $d_2 - 1$.

The conclusion is the same: SA articles, which are written by more informed investors and experts, reduce volatility of stock returns, whereas SA comments, which are posted mostly by individual investors, increase volatility of stock returns.

7 CONCLUSION

We use data of Seeking Alpha platform, retail order flow data, short sale flow data, and stock price data, to study the effect of social media on stock market trading. The main conclusion of the paper is that social media can facilitate informed trading. The tone of social media are informative in predicting retail order and short sale flows. Positive social media noises predicts more net buys by retail investors but more short sales by sophisticated investors. More information from social media let retail investors reduce contrarian trading. Social media articles make retail orders and short sale flows more informed. It is the bullish articles are most informative in increasing return predictability of retail order flows and short sale flows, and for both retail orders and short sales, social media are most helpful in identifying undervalued stocks. More publications of SA articles improve price efficiency, but more posts of SA comments slows the speed of impounding information into prices, which suggests that commentary on social media more reflects noise tradings.

In this paper we emphasize the role of investor tradings in understanding the interactions between social media coverage and stock price movements, and construct the portfolios that combine the signals of social media and signals of tradings (order flows of retail investors and short sellers) and achieve considerable risk-adjusted returns.

REFERENCES

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, 1259–1294.

Asquith, Paul, Parag A. Pathak, and Jay R. Ritter, 2005, Short interest, institutional ownership, and stock returns, *Journal of Financial Economics* 78, 243–276.

Avery, Christopher N., Judith A. Chevalier, and Richard J. Zeckhauser, 2016, The "CAPS" prediction system and stock market returns, *Review of Finance* 20, 1363–1381.

Baker, Malcolm, and Jeremy C. Stein, 2004, Market liquidity as a sentiment indicator, *Journal of Financial Markets* 7, 271–299.

Barber, Brad M., and Terrance Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *The Journal of Finance* 55, 773–806.

Barrot, Jean-Noel, Ron Kaniel, and David Sraer, 2016, Are retail traders compensated for providing liquidity?, *Journal of Financial Economics* 120, 146–168.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang, 2008, Which shorts are informed?, *The Journal of Finance* 63, 491–527.

Boehmer, Ekkehart, Charles M. Jones, Xiaoyan Zhang, and Xinran Zhang, 2021, Tracking retail investor activity, *The Journal of Finance* 76, 2249–2305.

Boehmer, Ekkehart, and Juan Wu, 2013, Short selling and the price discovery process, *The Review of Financial Studies* 26, 287–322.

Campbell, John L., Matthew D. DeAngelis, and James R. Moon, 2019, Skin in the game: Personal stock holdings and investors' response to stock analysis on social media, *Review of Accounting Studies* 24, 731–779.

Chen, Hailiang, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang, 2014, Wisdom of crowds: The value of stock opinions transmitted through social media, *The Review of Financial Studies* 27, 1367–1403.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *The Journal of Finance* 52, 1035–1058.

Das, Sanjiv R., and Mike Y. Chen, 2007, Yahoo! for amazon: Sentiment extraction from small talk on the web, *Management Science* 53, 1375–1388.

Diamond, Douglas W., and Robert E. Verrecchia, 1987, Constraints on short-selling and asset price adjustment to private information, *Journal of Financial Economics* 18, 277–311.

Dim, Chukwuma, 2021, Should retail investors listen to social media analysts? Evidence from text-implied beliefs, Working paper, Frankfurt School of Finance & Management, March 2021 Version.

Drake, Michael S., James R. Moon, Brady J. Twedt, and James D. Warren, 2022, Social media analysts and sell-side analyst research, *Review of Accounting Studies* 1–36.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg, 2012, How are shorts informed? Short sellers, news, and information processing, *Journal of Financial Economics* 105, 260–278.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of political economy* 81, 607–636.

Farrell, Michael, Clifton Green, Russell Jame, and Stanimir Markov, 2018, The democratization of investment research: Implications for retail investor profitability and firm liquidity, Working paper, August 2018 Version.

Gomez, Enrique, Frank Heflin, James Moon, and James Warren, 2020, Can financial analysis on social media help level the playing field among investors? Evidence from seeking alpha, Working Paper 18-45, Georgia Tech Scheller College of Business, June 2020 Version.

Grinblatt, Mark, and Matti Keloharju, 2001, What makes investors trade?, *The Journal of Finance* 56, 589–616.

Hapke, Hannes, Cole Howard, and Hobson Lane, 2019, *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python* (Simon and Schuster).

Hou, Kewei, and Tobias J. Moskowitz, 2005, Market frictions, price delay, and the cross-section of expected returns, *The Review of Financial Studies* 18, 981–1020.

Hu, Danqi, Charles M. Jones, Valerie Zhang, and Xiaoyan Zhang, 2021, The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery, Working paper, Available at SSRN 3807655.

Hvidkjaer, Soeren, 2008, Small trades and the cross-section of stock returns, *The Review of Financial Studies* 21, 1123–1151.

Kaniel, Ron, Shuming Liu, Gideon Saar, and Sheridan Titman, 2012, Individual investor trading and return patterns around earnings announcements, *The Journal of Finance* 67, 639–680.

Kaniel, Ron, Gideon Saar, and Sheridan Titman, 2008, Individual investor trading and stock returns, *The Journal of Finance* 63, 273–310.

Kelley, Eric K., and Paul C. Tetlock, 2013, How wise are crowds? Insights from retail orders and stock returns, *The Journal of Finance* 68, 1229–1265.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance* 66, 35–65.

McLean, R. David, Jeffrey Pontiff, and Christopher Reilly, 2020a, Retail investors and analysts, Working paper, October 2020 Version.

McLean, R. David, Jeffrey Pontiff, and Christopher Reilly, 2020b, Taking sides on return predictability, Working paper, Available at SSRN 3637649.

Odean, Terrance, 1999, Do investors trade too much?, *American Economic Review* 89, 1279–1298.

Senchack, Andrew J., and Laura T. Starks, 1993, Short-sale restrictions and market reaction to short-interest announcements, *Journal of Financial and Quantitative Analysis* 28, 177–194.

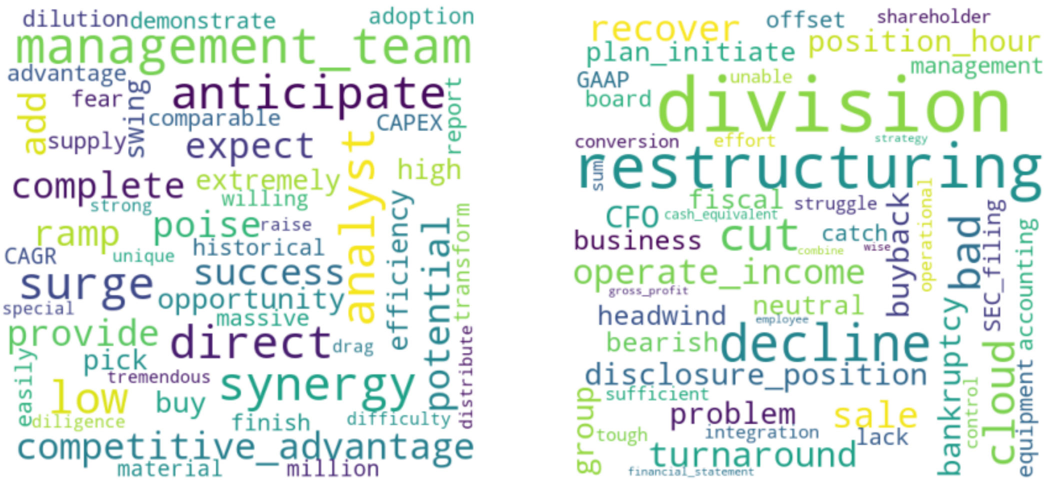
Shanthikumar, Devin, Annie Wang, and Shijia Wu, 2020, Social media and our opinions: How does social media interaction affect the extremeness of our opinions, Working paper, The Paul Merage School of Business, July 2020 Version.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *The Journal of Finance* 63, 1437–1467.

APPENDIX A FIGURES

Figure A.1: Words (Bigrams) Most Relevant to Retail Order Imbalances

This figure shows the top 50 words (bigrams) in SA articles most relevant to the retail order imbalances of next trading day (retail net buys and retail net sells respectively). The size of the word represents its relevance to the classification. To select the most relevant words (bigrams), we use the Supporting Vector Classifier method. For each article of a specific stock, we use the stock's order imbalance of next trading day as label (either net buy or net sell) and use TF-IDF of the words (bigrams) in the article as feature. We then use the labeled sample to train the model. We then use the weight vector of the trained model to identify the most relevant words for each classification.

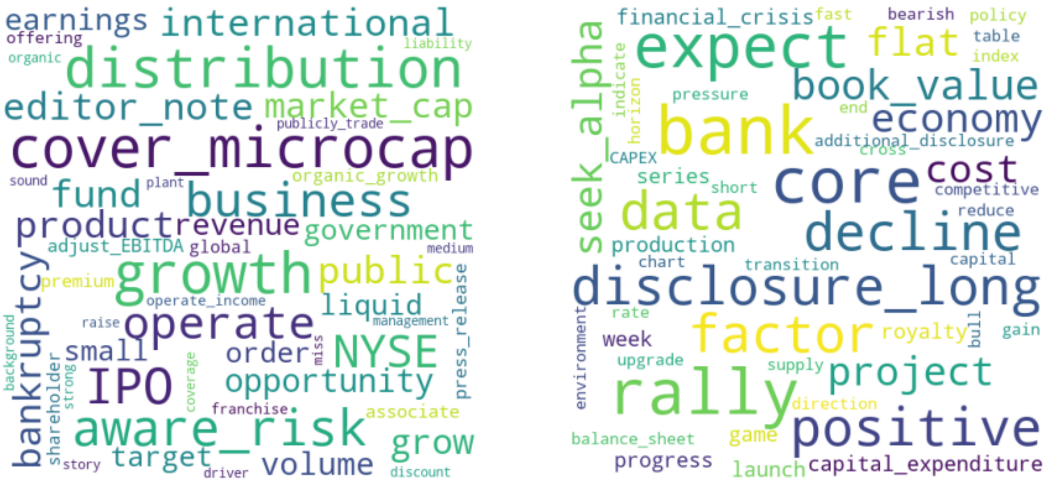


A. Retail net buys

B. Retail net sells

Figure A.2: Words (Bigrams) Most Relevant to Short Sale Flows

This figure shows the top 50 words (bigrams) in SA articles most relevant to the level of short sales of next trading day (lightly shorted and heavily shorted respectively). The size of the word represents its relevance to the classification. To select the most relevant words (bigrams), we use the Supporting Vector Classifier method. For each article of a specific stock, we use the stock's short sale level of next trading day as label (either lightly shorted or heavily shorted) and use TF-IDF of the words (bigrams) in the article as feature. We then use the labeled sample to train the model. We then use the weight vector of the trained model to identify the most relevant words for each classification.

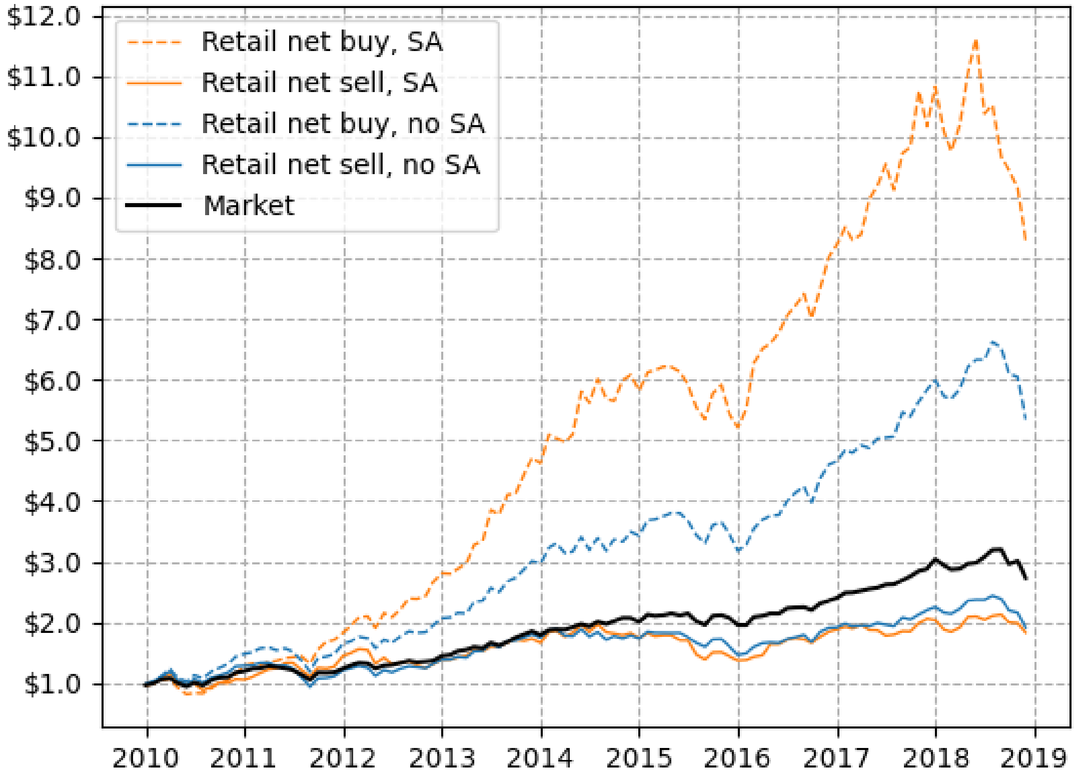


A. Lightly shorted

B. Heavily shorted

Figure A.3: Evolution of the Informed and Uninformed Retail Portfolios

This figure shows the monthly evolution of the $1 invested in the stocks that retail investors net buy and are covered by SA articles (the long leg of the informed portfolio), the stocks that retail investors net sell and are covered by SA articles (the short leg of the informed portfolio), the stocks that retail investors net buy but are not covered by SA articles (the long leg of the uninformed portfolio), the stocks that retail investors net sell but are not covered by SA articles (the short leg of the uninformed portfolio), and the market portfolio. The portfolios are rebalanced every trading day.

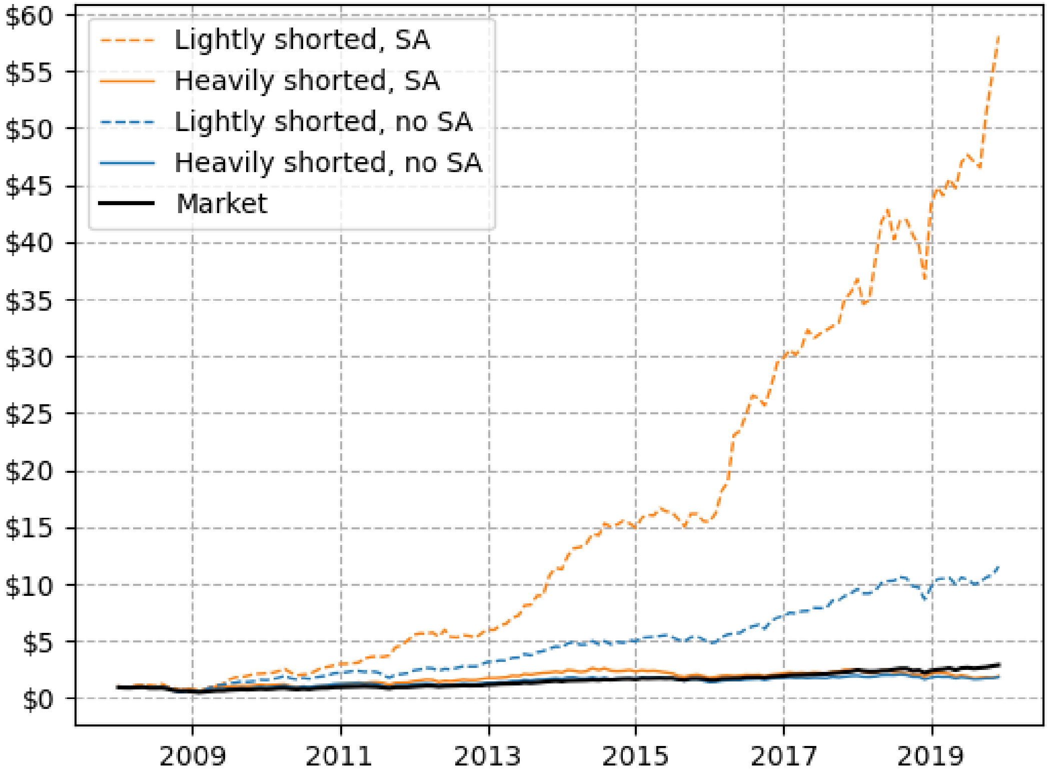


This table shows the mean of daily excess returns, standard deviation of daily returns, and Sharpe ratio of of the five portfolios in the figure above.

	Mean(excess return)	Std(return)	Sharpe ratio
Retail net buy, SA	0.0010	0.0134	0.08
Retail net sell, SA	0.0003	0.0125	0.03
Retail net buy, no SA	0.0008	0.0110	0.07
Retail net sell, no SA	0.0003	0.0108	0.03
Market	0.0005	0.0097	0.05

Figure A.4: Evolution of the Informed and Uninformed "Short Seller" Portfolios

This figure shows the monthly evolution of $1 invested in the stocks that are lightly shorted and covered by SA articles (the long leg of the informed portfolio), the stocks that are heavily shorted and covered by SA articles (the short leg of the informed portfolio), the stocks that are lightly shorted but not covered by SA articles (the long leg of the uninformed portfolio), the stocks that are heavily shorted but not covered by SA articles (the short leg of the uninformed portfolio), and the market portfolio. The portfolios are rebalanced every trading day.



This table shows the mean of daily excess returns, standard deviation of daily returns, and Sharpe ratio of the five portfolios in the figure above.

	Mean(excess return)	Std(return)	Sharpe ratio
Lightly shorted, SA	0.0019	0.0226	0.08
Heavily shorted, SA	0.0003	0.0167	0.02
Lightly shorted, no SA	0.0009	0.0137	0.06
Heavily shorted, no SA	0.0003	0.0148	0.02
Market	0.0004	0.0124	0.03

APPENDIX B TABLES

Table B.1: Descriptive Statistics

This table reports the descriptive statistics of major variables used in this paper. The sample period is from 2007 to 2019. The definitions of the variables are in Table C.1. Panel A reports statistics of individual variables. Panel B reports correlation between each of the two variables. The numbers in Panel A and B are time-series averages of the daily cross-sectional statistics. All the variables in Panel A and B are winsorized at 1% level. Panel C is annual statistics of the coverage of SA articles and comments of individual stocks.

Panel A: Summary statistics

Variable	Mean	Min	P25	Median	P75	Max	StdDev
Stock returns (Ret)	0.0006	-0.0874	-0.0123	-0.0002	0.012	0.1008	0.0374
Abnormal returns (AR)	0.0001	-0.0835	-0.0121	-0.0005	0.0111	0.0929	0.0345
SA Article tone ($\text{NegSA}^{\text{Article}}$)	0.033	0.003	0.019	0.03	0.044	0.093	0.021
SA comment tone ($\text{NegSA}^{\text{Comment}}$)	0.02	0	0.002	0.016	0.029	0.099	0.024
Retail net sell ($\text{NetSell}^{\text{vol}}$)	0.03	-1	-0.22	0.02	0.28	1	0.46
Retail net sell ($\text{NetSell}^{\text{trans}}$)	0.02	-1	-0.19	0.01	0.24	1	0.41
Short sales (SS)	0.08	0	0.04	0.07	0.1	0.24	0.05
Analyst coverage (Numest)	6.4	0	1.1	4.3	9.3	28.5	6.8
Analyst revision (Numup6)	1.2	0	0	0.2	1.4	11.8	2.4
Analyst revision (Numdown6)	1.8	0	0	0.6	2.3	14.9	3.1
News sentiment (ESS)	0.53	0.26	0.46	0.52	0.61	0.78	0.11
News coverage ($\log(\text{AEV})$)	1.82	0.39	1.3	1.72	2.22	4.13	0.75
Earnings growth ($\text{EarningsGrowth}^{1q}$)	0.002	-0.356	-0.007	0	0.007	0.393	1.18

Panel B: Correlation

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]
[1] Ret	1												
[2] AR	0.96	1											
[3] $\text{NegSA}^{\text{Article}}$	-0.07	-0.07	1										
[4] $\text{NegSA}^{\text{Comment}}$	-0.03	-0.03	0.17	1									
[5] $\text{NetSell}^{\text{vol}}$	-0.04	-0.04	0.02	0	1								
[6] $\text{NetSell}^{\text{trans}}$	-0.05	-0.05	0.03	0.01	0.82	1							
[7] SS	0.01	0.02	0.01	0.01	0.01	0	1						
[8] Numest	0	0	-0.05	0	-0.02	-0.02	0.2	1					
[9] Numup6	0.01	0.01	-0.06	-0.01	-0.01	-0.01	0.1	0.47	1				
[10] Numdown6	-0.01	-0.01	0.05	0.02	-0.01	-0.01	0.1	0.5	0.15	1			
[11] ESS	0.13	0.13	-0.11	-0.04	-0.01	-0.02	0	0.02	0.03	-0.02	1		
[12] $\log(\text{AEV})$	0	0.01	0.03	0.04	-0.01	-0.02	0.1	0.47	0.22	0.25	0.1	1	
[13] $\text{EarningsGrowth}^{1q}$	0	0	-0.02	0	0	0	0	0	0	0	0	0	1

Panel C: Annual SA coverage

Year	# of SA articles (1)	# of SA comments (2)	# of Stocks (3)	# of stocks Covered by SA articles (4)	# of stocks Covered by SA comments (5)
2007	10660	3775	5110	1451	505
2008	9355	10760	4797	1268	922
2009	10177	11336	4497	1195	926
2010	10120	13124	4281	1311	1042
2011	14002	23488	4108	1449	1269
2012	24311	47643	3980	1754	1601
2013	25855	61368	3911	2474	2289
2014	31463	74735	4000	2555	2468
2015	35675	87432	4037	2726	2621
2016	27936	86205	3952	2444	2453
2017	26961	94436	3885	2278	2450
2018	23942	83241	3892	2284	2378
2019	23001	76340	3892	2323	2401

Table B.2: Social Media Tone and Retail Order Imbalances

This table reports the results of regressing retail order imbalances of contemporaneous (t) or next trading day ($t+1$) on the average negativeness of SA articles and comments of stocks of day t. The definitions of variables are on table C.1. All continuous variables on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	By trading volume			By # of transactions		
	$\text{NetSell}_t^{\text{vol}}$	$\text{NetSell}_{t+1}^{\text{vol}}$	$\text{NetSell}_{t+1}^{\text{vol}}$	$\text{NetSell}_t^{\text{trans}}$	$\text{NetSell}_{t+1}^{\text{trans}}$	$\text{NetSell}_{t+1}^{\text{trans}}$
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{NegSA}_t^{\text{Article}}$	0.0043***	0.0031***	0.0027***	0.0061***	0.0047***	0.0036***
	(6.01)	(5.17)	(4.68)	(7.57)	(6.93)	(5.98)
$\text{NegSA}_t^{\text{Comment}}$	0.0008*	0.0010**	0.0010**	0.0019***	0.0015***	0.0011***
	(1.69)	(2.56)	(2.39)	(4.29)	(3.78)	(2.85)
AR_t	-0.0020***	-0.0022***	-0.0020***	-0.0034***	-0.0028***	-0.0022***
	(-3.08)	(-4.17)	(-3.94)	(-4.03)	(-4.76)	(-4.08)
$\text{AR}_{[t-5,t-1]}$	0.0045***	0.0034***	0.0030***	0.0045***	0.0041***	0.0033***
	(9.31)	(7.28)	(6.99)	(8.00)	(8.61)	(8.33)
Momentum	-0.0440***	-0.0558***	-0.0519***	-0.1305***	-0.1347***	-0.1103***
	(-2.77)	(-3.75)	(-3.77)	(-5.36)	(-6.80)	(-7.06)
Volatility	-0.0007	-0.0004	-0.0004	-0.0006	-0.0004	-0.0003
	(-1.25)	(-0.67)	(-0.62)	(-0.99)	(-0.69)	(-0.61)
$\text{NetSell}_t^{\text{vol}}$			0.0894***			
			(19.04)			
$\text{NetSell}_t^{\text{trans}}$						0.1861***
						(30.85)
$\text{I}_{[t-5,t-1]}^{\text{AR}}$	0.0038	0.0033	0.0029	0.0091**	0.0069**	0.0052**
	(1.15)	(1.28)	(1.27)	(2.20)	(2.18)	(2.11)
$\text{I}_t^{\text{Article}}$	-0.0107***	-0.0086***	-0.0077***	-0.0165***	-0.0159***	-0.0126***
	(-6.03)	(-5.94)	(-5.46)	(-7.37)	(-8.67)	(-8.16)
$\text{I}_t^{\text{Comment}}$	-0.0082***	-0.0081***	-0.0074***	-0.0201***	-0.0192***	-0.0156***
	(-3.11)	(-3.71)	(-3.55)	(-6.69)	(-7.63)	(-7.31)
$\text{I}^{\text{Momentum}}$	-0.0019	-0.0005	-0.0003	-0.0080	-0.0046	-0.0030
	(-0.43)	(-0.14)	(-0.10)	(-1.46)	(-1.01)	(-0.84)
$\text{I}_t^{\text{Retail}}$	0.0033**	0.0117	0.0113	-0.0082***	-0.0024	0.0027
	(2.02)	(0.85)	(0.83)	(-3.47)	(-0.20)	(0.23)
$\text{I}_{t+1}^{\text{Retail}}$		0.0050***	0.0051***			-0.0090***
		(3.03)	(3.09)			(-3.88)
$\text{I}^{\text{Volatility}}$	-0.0098	0.0000	0.0009	0.0102	-0.0009	-0.0019
	(-0.89)	(0.00)	(0.09)	(0.83)	(-0.09)	(-0.22)
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	385040	385040	385040	385040	385040	385040
Adj. R^2	0.0027	0.0025	0.0122	0.0072	0.0067	0.0494

Table B.3: The Impact of Social Media on Liquidity Provision of Retail Investors

This table reports the results of regressing net buys of retail investors of day $t+1$ on the interactions of abnormal returns and indicators of social media opinion access on prior periods $(t-p)$. The definitions of variables are on table C.1. All continuous variables on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	By trading volume			By # of transactions		
	$\mathrm{NetBuy}_{t+1}^{\mathrm{vol}}$	$\mathrm{NetBuy}_{t+1}^{\mathrm{vol}}$	$\mathrm{NetBuy}_{t+1}^{\mathrm{vol}}$	$\mathrm{NetBuy}_{t+1}^{\mathrm{trans}}$	$\mathrm{NetBuy}_{t+1}^{\mathrm{trans}}$	$\mathrm{NetBuy}_{t+1}^{\mathrm{trans}}$
	(1)	(2)	(3)	(4)	(5)	(6)
AR_t	-0.0009*	-0.0011**	-0.0011**	0.0008**	0.0007*	0.0007*
	(-1.97)	(-2.23)	(-2.23)	(2.04)	(1.75)	(1.76)
$\mathrm{AR}_{[t-5,\,t-1]}$	-0.0093***	-0.0097***	-0.0097***	-0.0076***	-0.0080***	-0.0080***
	(-14.07)	(-13.65)	(-13.65)	(-13.43)	(-12.87)	(-12.87)
$\mathrm{AR}_{[t-26,\,t-6]}$	-0.0073***	-0.0078***	-0.0078***	-0.0063***	-0.0067***	-0.0067***
	(-11.99)	(-11.67)	(-11.67)	(-10.42)	(-9.99)	(-9.99)
$\mathrm{AR}_t \times \mathrm{I}_t^{\mathrm{Articles}}$		0.0049***			0.0039***	
		(6.28)			(4.56)	
$\mathrm{AR}_{[t-5,\,t-1]} \times \mathrm{I}_{[t-5,\,t-1]}^{\mathrm{Articles}}$		0.0053***			0.0036***	
		(6.97)			(4.90)	
$\mathrm{AR}_{[t-26,\,t-6]} \times \mathrm{I}_{[t-26,\,t-6]}^{\mathrm{Articles}}$		0.0026***			0.0015**	
		(4.34)			(2.56)	
$\mathrm{AR}_t \times \mathrm{I}_t^{\mathrm{AC}}$			0.0048***			0.0037***
			(6.45)			(4.49)
$\mathrm{AR}_{[t-5,\,t-1]} \times \mathrm{I}_{[t-5,\,t-1]}^{\mathrm{AC}}$			0.0061***			0.0041***
			(7.64)			(5.19)
$\mathrm{AR}_{[t-26,\,t-6]} \times \mathrm{I}_{[t-26,\,t-6]}^{\mathrm{AC}}$			0.0032***			0.0019***
			(5.01)			(3.21)
$\mathrm{AR}_t \times \mathrm{I}_t^{\mathrm{A\overline{C}}}$			0.0053**			0.0054**
			(2.50)			(2.20)
$\mathrm{AR}_{[t-5,\,t-1]} \times \mathrm{I}_{[t-5,\,t-1]}^{\mathrm{A\overline{C}}}$			-0.0017			-0.0006
			(-1.01)			(-0.40)
$\mathrm{AR}_{[t-26,\,t-6]} \times \mathrm{I}_{[t-26,\,t-6]}^{\mathrm{A\overline{C}}}$			-0.0030			-0.0030
			(-1.51)			(-1.62)
$\mathrm{I}_t^{\mathrm{Articles}}$		0.0094***			0.0128***	
		(7.87)			(10.90)	
$\mathrm{I}_{[t-5,t-1]}^{\mathrm{Articles}}$		0.0048***			0.0084***	
		(6.10)			(10.23)	
$\mathrm{I}_{[t-26,t-6]}^{\mathrm{Articles}}$		0.0038***			0.0051***	
		(5.89)			(7.67)	
$\mathrm{I}_t^{\mathrm{AC}}$			0.0079***			0.0129***
			(6.33)			(9.32)
$\mathrm{I}_{[t-5,\,t-1]}^{\mathrm{AC}}$			0.0054***			0.0093***
			(5.98)			(9.46)
$\mathrm{I}_{[t-26,\,t-6]}^{\mathrm{AC}}$			0.0045***			0.0059***
			(5.91)			(7.46)
$\mathrm{I}_t^{\mathrm{A\overline{C}}}$			0.0126***			0.0121***
			(5.92)			(6.78)
$\mathrm{I}_{[t-5,\,t-1]}^{\mathrm{A\overline{C}}}$			0.0031**			0.0053***
			(2.11)			(3.87)
$\mathrm{I}_{[t-26,\,t-6]}^{\mathrm{A\overline{C}}}$			0.0015			0.0026**
			(1.09)			(2.03)
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	5651912	5651912	5651912	5651912	5651912	5651912
Adj. R^2	0.0083	0.0083	0.0083	0.012	0.0121	0.0121

Table B.4: The Impact of Social Media Coverage on Return Predictability of Retail Order Imbalances

This table reports the results of regressing abnormal returns of day t on interactions of retail order imbalance and indicators of social media coverage of day $t + 1$. The definitions of variables are on table C.1. All continuous variables on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	AR$_{t+1}$	AR$_{t+1}$	AR$_{t+1}$	AR$_{t+1}$
	(1)	(2)	(3)	(4)
NetBuy$_t$	0.0002***	0.0002***	0.0002***	
	(13.24)	(13.11)	(12.93)	
NetBuy$_t$ $\times$ I$_t^{\text{Article}}$		0.0005***		
		(3.21)		
NetBuy$_t$ $\times$ I$_t^{\text{AC}}$			0.0006**	
			(2.55)	
NetBuy$_t$ $\times$ I$_t^{\text{A}\overline{\text{C}}}$			0.0004*	
			(1.82)	
$\max$ (NetBuy$_t$, 0)				0.0008***
				(10.16)
$\max$ (NetBuy$_t$, 0) $\times$ I$_t^{\text{BullishArticle}}$				0.0031***
				(4.05)
$\max$ (NetSell$_t$, 0)				-0.0004***
				(-5.28)
$\max$ (NetSell$_t$, 0) $\times$ I$_t^{\text{BearishArticle}}$				0.0012
				(0.40)
AR$_t$	-0.0015***	-0.0015***	-0.0015***	-0.0017***
	(-12.14)	(-12.14)	(-12.14)	(-13.25)
AR$_{[t-5,t-1]}$	-0.0007***	-0.0007***	-0.0007***	-0.0007***
	(-9.58)	(-9.58)	(-9.57)	(-9.67)
AR$_{[t-26,t-6]}$	-0.0005***	-0.0005***	-0.0005***	-0.0005***
	(-9.58)	(-9.58)	(-9.57)	(-9.63)
I$_t^{\text{Article}}$		0.0000		
		(0.11)		
I$_t^{\text{AC}}$			-0.0001	
			(-0.47)	
I$_t^{\text{A}\overline{\text{C}}}$			0.0002	
			(1.02)	
I$_t^{\text{BearishArticle}}$				-0.0037***
				(-6.34)
I$_t^{\text{BullishArticle}}$				0.0001
				(1.02)
Month fixed effect	Yes	Yes	Yes	Yes
Firm fixed effect	Yes	Yes	Yes	Yes
Nobs	5651912	5651912	5651912	5651912
Adj. R^2	0.0035	0.0035	0.0035	0.004

Table B.5: Alpha and Risk Loading of Two Portfolios Mimicking Retail Investments

This table presents the annualized alphas and risk loadings of two zero-cost portfolios: the informed and uninformed portfolios. Both portfolios are rebalanced on a daily basis. Both portfolios are long the stocks that retail investors net buy at day t and are short the stocks that retail investors net sell at day t. The informed portfolio requires the stocks in both of its short leg and long leg have SA articles published at day t, the uninformed portfolio requires none of the stocks in either its short leg or long leg have SA articles published at day t. The LHS of the regressions are daily excess returns of the two portfolios on day $t + 1$. The daily factors on the RHS are Fama-French five factors, momentum factor, short term reversal factor, long term reversal factor on day $t + 1$.

	Informed		Uninformed	
	(1)	(2)	(3)	(4)
Alpha	0.1607***	0.1597***	0.1111***	0.1118***
	(2.96)	(2.94)	(15.29)	(15.36)
Mkt-rf	0.0515**	0.0717**	0.0030	0.0030
	(1.98)	(2.50)	(0.85)	(0.79)
SMB	-0.0335	-0.0285	0.0131**	0.0118*
	(-0.73)	(-0.62)	(2.13)	(1.90)
HML	-0.0555	-0.0374	-0.0059	-0.0119
	(-0.98)	(-0.60)	(-0.78)	(-1.42)
CMA	-0.0019	0.0365	-0.0046	-0.0089
	(-0.02)	(0.37)	(-0.38)	(-0.66)
RMW	-0.0211	-0.0530	0.0022	0.0048
	(-0.29)	(-0.68)	(0.22)	(0.46)
MOM		-0.0150		-0.0056
		(-0.43)		(-1.21)
STRev		-0.0548		-0.0034
		(-1.35)		(-0.63)
LTRev		-0.0797		0.0097
		(-1.06)		(0.96)
Nobs	2264	2264	2264	2264
Adj. R^2	0.0027	0.0041	0.004	0.0055

Table B.6: Social Media Tone and Short Sale Flows

This table reports the results of regressing Cboe short sales (adjusted by total CRSP trading volume) of contemporaneous trading day (t) or next trading day ($t+1$) on the average negativeness of SA articles and comments of stocks at date t. The definitions of variables are on table C.1. All continuous variables (except SS_t) on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	SS_t	SS_{t+1}	SS_{t+1}	SS_t	SS_{t+1}	SS_{t+1}
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{NegSA}_t^{\text{Article}}$	0.0005**	0.0005**	0.0003**	0.0004*	0.0004**	0.0002*
	(2.42)	(2.60)	(2.34)	(1.94)	(2.05)	(1.80)
$\text{NegSA}_t^{\text{Comment}}$				0.0002**	0.0002**	0.0001
				(2.15)	(2.02)	(1.27)
Volatility	-0.0005**	-0.0003**	-0.0001	-0.0005**	-0.0003**	-0.0001
	(-2.31)	(-2.18)	(-1.53)	(-2.33)	(-2.21)	(-1.62)
SS_t			0.4275***			0.4271***
			(64.49)			(64.79)
AR_t	0.0011***	0.0015***	0.0011***	0.0011***	0.0015***	0.0011***
	(8.39)	(10.13)	(9.92)	(8.41)	(10.12)	(9.90)
$\text{AR}_{[t-5,t-1]}$	0.0005***	0.0003***	0.0001**	0.0005***	0.0003***	0.0001**
	(3.91)	(3.61)	(2.30)	(3.99)	(3.71)	(2.42)
$\text{I}_t^{\text{Article}}$	-0.0014**	-0.0012**	-0.0005*	-0.0001	0.0001	0.0002
	(-2.20)	(-2.23)	(-1.82)	(-0.13)	(0.22)	(0.66)
$\text{I}_t^{\text{Comment}}$				0.0044***	0.0042***	0.0023***
				(4.71)	(5.68)	(6.19)
$\text{I}^{\text{Volatility}}$	0.0410***	0.0247***	0.0073***	0.0407***	0.0245***	0.0072***
	(12.15)	(9.43)	(4.70)	(12.17)	(9.43)	(4.65)
$\text{I}_{[t-5,t-1]}^{\text{AR}}$	0.0074***	0.0059***	0.0027***	0.0074***	0.0059***	0.0027***
	(4.58)	(4.63)	(4.52)	(4.59)	(4.64)	(4.54)
I_t^{SS}	0.0847***		-0.0298***	0.0843***		-0.0299***
	(26.55)		(-16.39)	(26.56)		(-16.50)
$\text{I}_{t+1}^{\text{SS}}$		0.0964***	0.0951***		0.0964***	0.0952***
		(81.85)	(80.84)		(81.78)	(80.81)
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	460136	460136	460136	460136	460136	460136
Adj. R^2	0.2238	0.5719	0.6657	0.2244	0.5723	0.6658

Table B.7: Projecting Social Media Tone on Informational Variables

This table reports the results of first step OLS regression of SA sentiment of articles and comments on proxy variables of information. We used the residuals of this regression as RHS variables in the step 2 regressions. The definitions of variables are on table C.1.

	$\text{NegSA}^{\text{Article}}$ (1)	$\text{NegSA}^{\text{Comment}}$ (2)
$\text{AdjEPS}_{\text{after}}$	0.0000	-0.0008***
	(0.01)	(-7.66)
$\text{AdjEPS}_{\text{before}}$	-0.0005***	-0.0008***
	(-4.39)	(-6.26)
$\text{EarningsGrowth}^{1q}_{\text{after}}$	0.0000	0.0000
	(0.27)	(0.90)
$\text{EarningsGrowth}^{1q}_{\text{before}}$	0.0000	0.0000
	(0.84)	(0.04)
$\text{EarningsGrowth}^{1y}_{\text{after}}$	0.0000	0.0000
	(0.14)	(0.14)
$\text{EarningsGrowth}^{1y}_{\text{before}}$	0.0000	0.0000
	(0.70)	(0.85)
Numdown1	0.0002***	0.0001***
	(13.28)	(5.18)
Numdown2	0.0000**	0.0000**
	(-2.22)	(-2.05)
Numdown3	0.0000	0.0003***
	(1.60)	(9.81)
Numdown4	0.0001**	-0.0002***
	(2.48)	(-3.12)
Numdown6	0.0001***	-0.0001***
	(3.25)	(-3.96)
Numdown7	0.0000	0.0000
	(-0.34)	(1.11)
Numdown8	0.0000	0.0000
	(-0.53)	(1.03)
Numdown9	0.0000	0.0000
	(0.60)	(0.81)
Numup1	0.0001***	0.0000
	(4.02)	(0.42)
Numup2	-0.0001***	-0.0001***
	(-3.80)	(-3.28)
Numup3	-0.0001**	0.0001**
	(-2.53)	(2.11)
Numup4	0.0002***	-0.0002**
	(3.41)	(-2.36)
Numup6	0.0001***	0.0000
	(4.45)	(-1.39)
Numup7	0.0000	0.0000
	(-0.16)	(0.82)
Numup8	0.0000**	0.0000
	(-2.18)	(1.55)

Table B.7 (continued)

Numup9	0.0000	0.0000
	(-1.01)	(0.22)
ESS	-0.0040***	-0.0064***
	(-11.17)	(-15.14)
Meanrec	0.0015***	0.0014***
	(23.80)	(17.86)
$\text{Rec}^{\text{Numdown}}$	0.0004***	0.0002***
	(9.56)	(3.57)
$\text{Rec}^{\text{Numup}}$	0.0006***	0.0000
	(9.86)	(0.02)
$\text{I}\left(\text{EarningsGrowth}^{1q}_{\text{after}}\right)$	-0.0001	0.0014***
	(-0.15)	(2.63)
$\text{I}\left(\text{EarningsGrowth}^{1q}_{\text{before}}\right)$	-0.0005	0.0004
	(-1.57)	(0.92)
$\text{I}\left(\text{EarningsGrowth}^{1y}_{\text{after}}\right)$	0.0003	0.0010***
	(0.93)	(2.68)
$\text{I}\left(\text{EarningsGrowth}^{1y}_{\text{before}}\right)$	0.0014***	-0.0009***
	(5.34)	(-2.96)
$\text{I}\left(\text{AdjEPS}_{\text{after}}\right)$	-0.0005*	-0.0037***
	(-1.67)	(-10.70)
$\text{I}\left(\text{AdjEPS}_{\text{before}}\right)$	-0.0007	0.0000
	(-1.48)	(-0.04)
I(Numdown1)	-0.0012***	0.0003
	(-4.97)	(0.96)
I(Numdown2)	0.0000	-0.0004**
	(0.00)	(-2.18)
I(Numdown3)	0.0012***	-0.0008***
	(12.23)	(-6.44)
I(Numdown4)	0.0011***	0.0003***
	(12.52)	(3.19)
I(Numdown6)	-0.0008**	-0.0007**
	(-2.52)	(-2.03)
I(Numdown7)	0.0005*	0.0000
	(1.65)	(0.14)
I(Numdown8)	0.0001	0.0006**
	(0.49)	(2.35)
I(Numdown9)	0.0006***	-0.0012***
	(3.54)	(-6.24)
I(Ravenpack)	0.0043***	0.0039***
	(21.37)	(16.31)
$\text{I}(\text{Rec}^{\text{Numdown}})$	-0.0013***	-0.0018***
	(-4.37)	(-4.86)
(Intercept)	0.0073***	0.0186***
	(43.75)	(93.06)
Nobs	460136	460136
Adj. R^2	0.0198	0.004

Table B.8: Social Media Noises and Short Sale Flows

This table shows the regression of short sales of contemporaneous trading day t or next trading day $t + 1$ on abnormal social media sentiment (the part of sentiment that cannot be explained by information) at day t. All continuous variables (except SS_t) on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The definitions of variables are on table C.1. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	SS_t	SS_{t+1}	SS_{t+1}	SS_t	SS_{t+1}	SS_{t+1}
	(1)	(2)	(3)	(4)	(5)	(6)
$\mathrm{AbNegSA}^{\mathrm{Article}}$	-0.0012***	-0.0008***	-0.0001**	-0.0011***	-0.0009***	-0.0003***
	(-4.43)	(-3.62)	(-2.16)	(-4.17)	(-4.15)	(-2.97)
$\mathrm{AbNegSA}^{\mathrm{Comment}}$				0.0001*	0.0001	0.0001
				(1.77)	(1.21)	(1.01)
Volatility	-0.0005**	-0.0003**	-0.0001	-0.0005**	-0.0003**	-0.0001
	(-2.23)	(-2.11)	(-1.46)	(-2.28)	(-2.22)	(-1.51)
SS_t			0.4275***			0.4270***
			(64.56)			(64.87)
AR_t	0.0010***	0.0015***	0.0011***	0.0010***		0.0011***
	(8.36)	(10.10)	(9.90)	(8.30)		(9.88)
$\mathrm{AR}_{[t-5,t-1]}$	0.0005***	0.0003***	0.0001**	0.0005***	0.0003***	0.0001**
	(3.83)	(3.40)	(2.17)	(3.81)	(3.48)	(2.24)
$\mathrm{I}_t^{\mathrm{Article}}$	0.0015*	0.0009		0.0024***	0.0022***	0.0011***
	(1.94)	(1.50)		(2.63)	(3.03)	(3.11)
$\mathrm{I}_t^{\mathrm{Comment}}$				0.0045***	0.0044***	0.0024***
				(4.88)	(5.86)	(6.35)
$\mathrm{I}^{\mathrm{Volatility}}$	0.0417***	0.0248***	0.0073***	0.0414***	0.0246***	0.0072***
	(12.14)	(9.49)	(4.71)	(14.56)	(10.03)	(4.69)
$\mathrm{I}^{\mathrm{AR}}_{[t-5,t-1]}$	0.0073***	0.0058***	0.0027***	-0.0051	0.0037	0.0027***
	(4.49)	(4.58)	(4.51)	(-0.58)	(0.89)	(4.50)
$\mathrm{I}_t^{\mathrm{SS}}$			-0.0298***	0.0839***		-0.0300***
			(-16.39)	(26.10)		(-16.50)
$\mathrm{I}_{t+1}^{\mathrm{SS}}$		0.0964***	0.0951***		0.0965***	0.0952***
		(81.83)	(80.79)		(81.68)	(80.80)
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	460136	460136	460136	460136	460136	460136
Adj. R^2	0.2059	0.5719	0.6657	0.2247	0.5716	0.6658

Table B.9: Impact of Social Media on Return Predictability of Short Sales

This table presents the results of regressing abnormal return of next trading day on the interactions of adjusted short sales and indicators of social media coverage. All continuous variables (except SS_t) on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The definitions of variables are on table C.1. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	AR_{t+1}	AR_{t+1}	AR_{t+1}	AR_{t+1}
	(1)	(2)	(3)	(4)
SS_t	-0.0002***	-0.0002***	-0.0002***	-0.0002***
	(-7.96)	(-7.82)	(-7.81)	(-7.82)
$\mathrm{SS}_t \times \mathrm{I}_t^{\mathrm{Article}}$		-0.0005***		
		(-3.99)		
$\mathrm{SS}_t \times \mathrm{I}_t^{\mathrm{AC}}$			-0.0006***	
			(-3.34)	
$\mathrm{SS}_t \times \mathrm{I}_t^{\mathrm{A\overline{C}}}$			-0.0004*	
			(-1.84)	
$\mathrm{SS}_t \times \mathrm{I}_t^{\mathrm{BullishArticle}}$				-0.0005***
				(-3.73)
$\mathrm{SS}_t \times \mathrm{I}_t^{\mathrm{BearishArticle}}$				-0.0006
				(-0.79)
AR_t	-0.0017***	-0.0017***	-0.0017***	-0.0017***
	(-12.83)	(-12.83)	(-12.83)	(-12.83)
$\mathrm{AR}_{[t-5,t-1]}$	-0.0008***	-0.0008***	-0.0008***	-0.0008***
	(-8.44)	(-8.43)	(-8.44)	(-8.43)
$\mathrm{AR}_{[t-26,t-6]}$	-0.0004***	-0.0004***	-0.0004***	-0.0004***
	(-5.60)	(-5.60)	(-5.60)	(-5.60)
$\mathrm{I}_t^{\mathrm{Article}}$		0.0003**		
		(2.49)		
$\mathrm{I}_t^{\mathrm{AC}}$			0.0003**	
			(2.12)	
$\mathrm{I}_t^{\mathrm{A\overline{C}}}$			0.0002	
			(1.42)	
$\mathrm{I}_t^{\mathrm{BullishArticle}}$				0.0006***
				(5.36)
$\mathrm{I}_t^{\mathrm{BearishArticle}}$				-0.0029***
				(-5.08)
Firm fixed effect	Yes	Yes	Yes	Yes
Month fixed effect	Yes	Yes	Yes	Yes
Nobs	9525486	9525486	9525486	9525486
Adj. R^2	0.0033	0.0033	0.0033	0.0034

Table B.10: Alphas and Risk Loadings of Two portfolios Based on Short Sale Flow

This table presents risk loadings and annualized alphas of two zero-cost portfolios: informed and uninformed. Both portfolios are rebalanced on a daily basis. For each trading day t, we sort the stocks into two groups by the midpoint of adjusted short sale SS_t. Both portfolios are long the stocks in the low SS_t group and are short the stocks in the high SS_t group. The informed portfolio requires the stocks in both its long and short legs have SA articles published at day t, the uninformed portfolio requires none of the stocks in either its short leg or long leg have SA articles published at day t. The LHS of the regressions are daily excess returns on day $t + 1$. The monthly factors on the RHS are Fama-French five factors, momentum factor, short term reversal factor, and long term reversal factor. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	Informed		Uninformed	
	(1)	(2)	(3)	(4)
Alpha	0.3835***	0.3820***	0.1450***	0.1432***
	(4.29)	(4.27)	(9.44)	(9.34)
Mkt-rf	0.0416	0.0319	0.0062	-0.0008
	(1.24)	(0.90)	(1.07)	(-0.14)
SMB	0.0532	0.0643	0.0149	0.0212*
	(0.83)	(0.97)	(1.35)	(1.86)
HML	0.0229	-0.0161	0.0392***	0.0319**
	(0.38)	(-0.22)	(3.80)	(2.51)
CMA	-0.0885	0.0381	-0.1003***	-0.0595**
	(-0.72)	(0.27)	(-4.77)	(-2.49)
RMW	-0.0612	-0.1331	-0.0459**	-0.0629***
	(-0.57)	(-1.16)	(-2.50)	(-3.19)
MOM		-0.0798*		-0.0202***
		(-1.77)		(-2.61)
STRev		-0.0124		0.0152**
		(-0.28)		(2.03)
LTRev		-0.1292		-0.0393**
		(-1.44)		(-2.55)
Nobs	3014	3014	3014	3014
Adj. R^2	0.0023	0.004	0.0236	0.0293

Table B.11: The Impact of Social Media on Informational Efficiency of Stock Prices

This table reports the regression of price-delay measures of stocks in month t on the contemporaneous log number of articles and comments published about that stock. The definitions of variables are on table C.1. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	$D_{1,t}$	$D_{1,t}$	$D_{2,t}$	$D_{2,t}$	$D_{3,t}$	$D_{3,t}$
	(1)	(2)	(3)	(4)	(5)	(6)
$\log(\#)_t^{\text{Article}}$	-0.0197***	-0.0161***	-0.0293***	-0.0238***	-0.0291***	-0.0237***
	(-5.39)	(-4.42)	(-3.45)	(-2.82)	(-3.42)	(-2.82)
$\log(\#)_t^{\text{Comment}}$	0.0093***	0.0116***	0.0146***	0.0185***	0.0148***	0.0187***
	(3.94)	(5.05)	(2.80)	(3.60)	(2.87)	(3.68)
$D_{1,t-1}$	0.1143***	0.1060***				
	(14.72)	(14.05)				
$D_{2,t-1}$			0.0447***	0.0403***		
			(7.94)	(7.34)		
$D_{3,t-1}$					0.0447***	0.0403***
					(7.98)	(7.40)
Size		-0.0716***		-0.1259***		-0.1241***
		(-8.94)		(-8.24)		(-8.16)
B/M		-0.0008**		-0.0018**		-0.0017**
		(-2.04)		(-2.21)		(-2.05)
Momentum		-0.0200***		-0.0254**		-0.0250**
		(-4.10)		(-2.59)		(-2.61)
Analyst coverage		-0.0494***		-0.0854***		-0.0861***
		(-7.29)		(-6.22)		(-6.25)
log(AEV)		-0.0076***		-0.0113***		-0.0110***
		(-4.01)		(-3.16)		(-3.00)
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	294918	294918	294918	294918	294918	294918
Adj. R^2	0.4614	0.4659	0.3313	0.3344	0.3183	0.3214

APPENDIX C SUPPLEMENTARY TABLES

Table C.1: List of Variable Definitions

Variable	Definition
$\text{I}_t^{\text{Article}}$	Dummy variable indicating that there is SA article about the stock at day t.
$\text{I}_t^{\text{Comment}}$	Dummy variable indicating that there is SA comment about the stock at day t.
$\log(\#)_t^{\text{Article}}$	Log number of SA aritlces of a stock in a month or day.
$\log(\#)_t^{\text{Comment}}$	Log number of SA comments of a stock in a month or day.
$\text{NetSell}_t^{\text{vol}}$	Sell volume of retail investors minus buy volume of retail investors, divided by total volume
$\text{NetBuy}_t^{\text{vol}}$	$-\text{NetSell}_t^{\text{vol}}$
$\text{NetSell}_t^{\text{trans}}$	# of Sell transactions of retail investors minus # of buy transactions of retail investors, divided by # of total transactions
$\text{NetBuy}_t^{\text{trans}}$	- Net_sell_trans
$\text{NegSA}_t^{\text{Article}}$	Negativeness of Seekingalpha articles. It is defined as the daily average of the negativeness of articles about a stock of day t. Negativeness of an article is defined as the fraction of negative words of the article matched with the Loughran-McDonald negative word list.
$\text{NegSA}_t^{\text{Comment}}$	Negativeness of Seekingalpha comments. It is defined as the daily average of the negativeness of articles about a stock of day t. Negativeness of an article is defined as the fraction of negative words of the article matched with the Loughran-McDonald negative word list.
$\text{AbNegSA}^{\text{Article}}$	The part of negativeness of SA article that is not explained by information of stock fundamentals. It is computed as the residuals from the regressions of negativeness of SA article on a group of variables that represent information of the stock.
$\text{AbNegSA}^{\text{Comment}}$	The part of negativeness of SA comment that is not explained by information of stock fundamentals. It is computed as the residuals from the regressions of negativeness of SA comment on a group of variables that represent information of the stock.
ESS	Ravenpack news event sentiment (ESS) divided by 100.
$\log(\text{AEV})$	log value of Ravenpack aggregate event volume.
Volatility	The variance of stock returns of prior 21 trading days.
Momentum	Return of prior year of date t of a stock.
SS_t	The ratio of short selling volume divided by total trading volume. Data is from CBOE short sale daily report.
AR_t	DGTW adjusted returns at day t.
$\text{AR}_{[t-5,t-1]}$	DGTW adjusted cumulative returns from day t-5 to day t-1.
$\text{AdjEPS}_{\text{after}}$	Earnings per share (adjusted by the closing price of the announcement day) announced after and nearest to the publishing day of SA article (comment).

Table C.1 (continued)

Variable	Definition
$\text{AdjEPS}_{\text{before}}$	Earnings per share (adjusted by the closing price of the announcement day) announced before and nearest to the publishing day of SA article (comment).
$\text{EarningsGrowth}^{1q}_{\text{after}}$	Earnings growth (compared to that of 1 quarter ago) computed using EPS announced after and nearest to the publishing day of SA article (comment).
$\text{EarningsGrowth}^{1q}_{\text{before}}$	Earnings growth (compared to that of 1 quarter ago) computed using EPS announced before and nearest to the publishing day of SA article (comment).
$\text{EarningsGrowth}^{1y}_{\text{after}}$	Earnings growth (compared to that of 1 year ago) computed using EPS announced after and nearest to the publishing day of SA article (comment).
$\text{EarningsGrowth}^{1y}_{\text{before}}$	Earnings growth (compared to that of 1 year ago) computed using EPS announced before and nearest to the publishing day of SA article (comment).
Numdown1	Number of downward analyst forecast revisions of the EPS of current year. The revisions are made in the month of date t.
Numdown2	Number of downward analyst forecast revisions of the EPS of next year. The revisions are made in the month of date t.
Numdown3	Number of downward analyst forecast revisions of the EPS two years later. The revisions are made in the month of date t.
Numdown4	Number of downward analyst forecast revisions of the EPS three years later. The revisions are made in the month of date t.
Numdown6	Number of downward analyst forecast revisions of the EPS of current quarter. The revisions are made in the month of date t.
Numdown7	Number of downward analyst forecast revisions of the EPS of next quarter. The revisions are made in the month of date t.
Numdown8	Number of downward analyst forecast revisions of the EPS two quarters later. The revisions are made in the month of date t.
Numdown9	Number of downward analyst forecast revisions of the EPS three quarters later. The revisions are made in the month of date t.
Numup1	Number of upward analyst forecast revisions of the EPS of current year. The revisions are made in the month of date t.
Numup2	Number of upward analyst forecast revisions of the EPS of next year. The revisions are made in the month of date t.
Numup3	Number of upward analyst forecast revisions of the EPS two years later. The revisions are made in the month of date t.
Numup4	Number of upward analyst forecast revisions of the EPS three years later. The revisions are made in the month of date t.
Numup6	Number of upward analyst forecast revisions of the EPS of current quarter. The revisions are made in the month of date t.
Numup7	Number of upward analyst forecast revisions of the EPS of next quarter. The revisions are made in the month of date t.
Numup8	Number of upward analyst forecast revisions of the EPS two quarters later. The revisions are made in the month of date t.

Table C.1 (continued)

Variable	Definition
Numup9	Number of upward analyst forecast revisions of the EPS three quarters later. The revisions are made in the month of date t.
Meanrec	Mean of IBES analyst recommendation scores. The recommendation scores are: 1 (Strong buy), 2 (Buy), 3 (Hold), 4 (Sell), 5 (Strong sell)
$\text{Rec}^{\text{Numdown}}$	Number of downward IBES analyst recommendation revisions.
$\text{Rec}^{\text{Numup}}$	Number of upward IBES analyst recommendation revisions
$\text{I}^{\text{Volatility}}$	Dummy varialbe indicating that the volatility variable is not null.
$\text{I}^{\text{AR}}_{[t-5,t-1]}$	Dummy variable indicating that the $\text{AR}_{[t-5,t-1]}$ variable is not null.
I^{SS}_{t}	Dummy variable indicating that the SS_t variable is not missing.
$\text{I}\left(\text{EarningsGrowth}^{1q}_{\text{after}}\right)$	Dummy variable indicating that the variable $\text{EarningsGrowth}^{1q}_{\text{after}}$ is not null.
$\text{I}\left(\text{EarningsGrowth}^{1q}_{\text{before}}\right)$	Dummy variable indicating that the variable $\text{EarningsGrowth}^{1q}_{\text{before}}$ is not null.
$\text{I}\left(\text{EarningsGrowth}^{1y}_{\text{after}}\right)$	Dummy variable indicating that the variable $\text{EarningsGrowth}^{1y}_{\text{after}}$ is not null.
$\text{I}\left(\text{EarningsGrowth}^{1y}_{\text{before}}\right)$	Dummy variable indicating that the variable $\text{EarningsGrowth}^{1y}_{\text{before}}$ is not null.
$\text{I}\left(\text{AdjEPS}^{1q}_{\text{after}}\right)$	Dummy variable indicating that the variable $\text{AdjEPS}^{1q}_{\text{after}}$ is not null.
$\text{I}\left(\text{AdjEPS}^{1q}_{\text{before}}\right)$	Dummy variable indicating that the variable $\text{AdjEPS}^{1q}_{\text{before}}$ is not null.
I(Numdown1)	Dummy variable indicating that the variable Numdown1 is not null.
I(Numdown2)	Dummy variable indicating that the variable Numdown2 is not null.
I(Numdown3)	Dummy variable indicating that the variable Numdown3 is not null.
I(Numdown4)	Dummy variable indicating that the variable Numdown4 is not null.
I(Numdown6)	Dummy variable indicating that the variable Numdown6 is not null.
I(Numdown7)	Dummy variable indicating that the variable Numdown7 is not null.
I(Numdown8)	Dummy variable indicating that the variable Numdown8 is not null.
I(Numdown9)	Dummy variable indicating that the variable Numdown9 is not null.
$\text{I}\left(\text{Rec}^{\text{Down}}\right)$	Dummy variable indicating that the variable Rec^{Down} is not null.
I(Ravenpack)	Dummy variable indicating that there is Ravenpack news about the stock at day t.

Table C.2: Social Media Noises and Retail Order Imbalances

This table shows the regression of net sell by retail investors on abnormal social media sentiment (the part of sentiment that cannot be explained by information). All continuous variables on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The definitions of variables are on table C.1. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	By trading volume			By # of transactions		
	$\text{NetSell}_t^{\text{vol}}$	$\text{NetSell}_{t+1}^{\text{vol}}$	$\text{NetSell}_{t+1}^{\text{vol}}$	$\text{NetSell}_t^{\text{trans}}$	$\text{NetSell}_{t+1}^{\text{trans}}$	$\text{NetSell}_{t+1}^{\text{trans}}$
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{AbNegSA}_t^{\text{Article}}$	0.0041***	0.0034***	0.0030***	0.0064***	0.0055***	0.0043***
	(5.55)	(6.17)	(5.73)	(7.22)	(7.81)	(7.09)
$\text{AbNegSA}_t^{\text{Comment}}$	0.0006	0.0009**	0.0008*	0.0016***	0.0013***	0.0010**
	(1.22)	(2.09)	(1.97)	(3.68)	(3.17)	(2.41)
AR_t	-0.0021***	-0.0022***	-0.0020***	-0.0035***	-0.0028***	-0.0022***
	(-3.12)	(-4.20)	(-3.97)	(-4.06)	(-4.76)	(-4.11)
$\text{AR}_{[t-5,t-1]}$	0.0045***	0.0034***	0.0030***	0.0045***	0.0041***	0.0032***
	(9.28)	(7.25)	(6.96)	(7.96)	(8.56)	(8.28)
Momentum	-0.0448***	-0.0563***	-0.0523***	-0.1316***	-0.1349***	-0.1105***
	(-2.84)	(-3.78)	(-3.79)	(-5.41)	(-6.81)	(-7.06)
Volatility	-0.0007	-0.0004	-0.0004	-0.0006	-0.0004	-0.0003
	(-1.28)	(-0.70)	(-0.64)	(-1.03)	(-0.72)	(-0.64)
$\text{NetSell}_t^{\text{vol}}$			0.0894***			
			(19.04)			
$\text{NetSell}_t^{\text{trans}}$						0.1860***
						(30.84)
$\text{I}_{[t-5,t-1]}^{\text{AR}}$	0.0039	0.0034	0.0031	0.0094**	0.0072**	0.0054**
	(1.20)	(1.34)	(1.32)	(2.27)	(2.25)	(2.20)
$\text{I}_t^{\text{Article}}$	-0.0102***	-0.0090***	-0.0081***	-0.0168***	-0.0167***	-0.0136***
	(-5.34)	(-6.00)	(-5.65)	(-6.60)	(-8.22)	(-8.07)
$\text{I}_t^{\text{Comment}}$	-0.0078***	-0.0078***	-0.0071***	-0.0195***	-0.0189***	-0.0153***
	(-2.95)	(-3.57)	(-3.42)	(-6.53)	(-7.48)	(-7.17)
$\text{I}^{\text{Momentum}}$	-0.0015	-0.0002	0.0000	-0.0074	-0.0040	-0.0026
	(-0.34)	(-0.05)	(-0.01)	(-1.35)	(-0.89)	(-0.73)
$\text{I}^{\text{Volatility}}$	-0.0094	0.0003	0.0011	0.0107	0.0004	-0.0016
	(-0.86)	(0.03)	(0.12)	(0.88)	(0.04)	(-0.18)
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	385040	385040	385040	385040	385040	385040
Adj. R^2	0.0027	0.0025	0.0123	0.0073	0.007	0.0494

Table C.3: Return Predictability of Social Media Tone

This table reports the results of regressing the cumulative DGTW-adjusted returns of next 5 trading days [t+1, t+5] (or skipping one day, [t+2, t+6]) on the average tones of Seekingalpha articles and comments at date t. The definitions of variables are on table C.1. All continuous variables on the RHS are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	$\text{AR}_{[t+1,t+5]}$	$\text{AR}_{[t+1,t+5]}$	$\text{AR}_{[t+1,t+5]}$	$\text{AR}_{[t+2,t+6]}$	$\text{AR}_{[t+2,t+6]}$	$\text{AR}_{[t+2,t+6]}$
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{NegSA}_t^{\text{Article}}$	-0.0006***	-0.0006***	-0.0005**	-0.0003	-0.0002	-0.0002
	(-2.96)	(-2.80)	(-2.29)	(-1.48)	(-1.32)	(-1.20)
$\text{NegSA}_t^{\text{Comment}}$		-0.0003**	-0.0002**		0.0000	0.0000
		(-2.24)	(-2.00)		(-0.42)	(-0.37)
ESS			0.0042***			0.0009**
			(9.92)			(2.24)
Volatility	-0.0002	-0.0002	-0.0002	-0.0001	-0.0001	-0.0001
	(-0.56)	(-0.55)	(-0.50)	(-0.27)	(-0.26)	(-0.24)
AR_t	-0.0008***	-0.0008***	-0.0010***	-0.0007***	-0.0007***	-0.0007***
	(-2.89)	(-2.92)	(-3.43)	(-2.66)	(-2.66)	(-2.77)
$\text{AR}_{[t-5,t-1]}$	-0.0012***	-0.0012***	-0.0012***	-0.0011***	-0.0011***	-0.0011***
	(-2.75)	(-2.77)	(-2.87)	(-2.77)	(-2.78)	(-2.81)
$\text{I}_t^{\text{Article}}$	0.0016***	0.0013***	0.0011***	0.0009**	0.0006	0.0005
	(3.88)	(3.11)	(2.65)	(2.50)	(1.52)	(1.32)
$\text{I}_t^{\text{Comment}}$		-0.0009*	-0.0009*		-0.0012***	-0.0013***
		(-1.85)	(-1.93)		(-2.83)	(-2.94)
I(Ravenpack)			-0.0084***			-0.0015*
			(-9.84)			(-1.89)
$\text{I}^{\text{Volatility}}$	-0.0004	-0.0003	-0.0001	-0.0007	-0.0006	-0.0006
	(-1.51)	(-1.19)	(-0.18)	(-1.16)	(-1.03)	(-1.00)
$\text{I}^{\text{AR}}_{[t-5,t-1]}$	-0.0001	-0.0001	0.0000	0.0000	0.0000	0.0000
	(-0.29)	(-0.28)	(-0.16)	(-0.02)	(-0.01)	(-0.14)
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	460136	460136	460136	460138	460138	460138
Adj. R^2	0.0017	0.0018	0.0021	0.0016	0.0017	0.0017

Table C.4: The Impact of Social Media on Informational Efficiency of Stocks (Binary Variables on RHS)

This table reports the regression of price-delay measures of month t on the dummy variables $\text{I}^{\text{Article}}$ and $\text{I}^{\text{Comment}}$ which indicate whether there are articles or comments from SA on month t. The definitions of variables are on table C.1. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	$\text{D}_{1,t}$	$\text{D}_{1,t}$	$\text{D}_{2,t}$	$\text{D}_{2,t}$	$\text{D}_{3,t}$	$\text{D}_{3,t}$
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{I}_t^{\text{Article}}$	-0.0140***	-0.0105***	-0.0208***	-0.0152**	-0.0206***	-0.0151**
	(-5.09)	(-3.88)	(-3.40)	(-2.53)	(-3.35)	(-2.51)
$\text{I}_t^{\text{Comment}}$	0.0074**	0.0117***	0.0167**	0.0239***	0.0172**	0.0243***
	(2.11)	(3.50)	(2.17)	(3.22)	(2.23)	(3.28)
	0.1143***	0.1061***				
$\text{D}_{1,t-1}$	(14.74)	(14.07)				
$\text{D}_{2,t-1}$			0.0446***	0.0402***		
			(7.94)	(7.35)		
$\text{D}_{3,t-1}$					0.0447***	0.0403***
					(7.99)	(7.41)
Size		-0.0713***		-0.1256***		-0.1237***
		(-8.92)		(-8.24)		(-8.16)
B/M		-0.0008**		-0.0018**		-0.0017**
		(-2.03)		(-2.21)		(-2.05)
Momentum		-0.0200***		-0.0254**		-0.0250**
		(-4.10)		(-2.58)		(-2.60)
Analyst coverage		-0.0491***		-0.0851***		-0.0857***
		(-7.24)		(-6.19)		(-6.22)
log AEV		-0.0076***		-0.0114***		-0.0110***
		(-4.00)		(-3.17)		(-3.01)
Month fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Firm fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Nobs	294918	294918	294918	294918	294918	294918
Adj. R^2	0.4614	0.4658	0.3313	0.3344	0.3183	0.3214

Table C.5: The Impact of Social Media on Stock Price Volatility

This table reports the results of Fama-MacBeth regression of sbusequent month ($[t+1, t+21]$) volatility of a stock on the number of SA articles and comments at t (column 1 and 2) or on indicators of publications of SA articles and comments at t (column 3 and 4). The Fama-MacBeth procedure is of two steps: we first run cross-sectional regressions for each date t; we then compute the coefficients and t value using the time series of the estimated parameters from step 1. To adjust for overlapping of the volatility variable, we compute Newey-West standard errors for 30 lags. The definitions of variables are on table C.1. The superscripts *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	Volatility$_{[t+1,t+21]}$	Volatility$_{[t+1,t+21]}$	Volatility$_{[t+1,t+21]}$	Volatility$_{[t+1,t+21]}$
	(1)	(2)	(3)	(4)
$\log(\#)_t^{\text{Article}}$	-0.0010***	-0.0007***		
	(-4.05)	(-2.87)		
$\log(\#)_t^{\text{Comment}}$	0.0005***	0.0002**		
	(4.42)	(2.28)		
$\text{I}_t^{\text{Article}}$			-0.0003***	-0.0002**
			(-2.69)	(-2.41)
$\text{I}_t^{\text{Comment}}$			0.0003**	0.0002
			(2.44)	(1.47)
log(AEV)	-0.0004***	-0.0003***	-0.0004***	-0.0003***
	(-14.47)	(-13.50)	(-14.04)	(-13.05)
Volatility$_{[t-21,t-1]}$	0.2574***	0.2814***	0.2574***	0.2815***
	(8.23)	(8.62)	(8.23)	(8.62)
AR$_t$		-0.0024**		-0.0024**
		(-2.42)		(-2.43)
AR$_{[t-5,t-1]}$		-0.0063***		-0.0063***
		(-6.49)		(-6.50)
AR$_{[t-26,t-6]}$		-0.0046***		-0.0046***
		(-8.89)		(-8.89)
(Intercept)	0.0016***	0.0122***	0.0016***	0.0122***
	(15.23)	(8.78)	(15.23)	(8.79)
Nobs	12622507	12589433	12622507	12589433
Avg. R^2	0.0461	0.0671	0.046	0.0671